

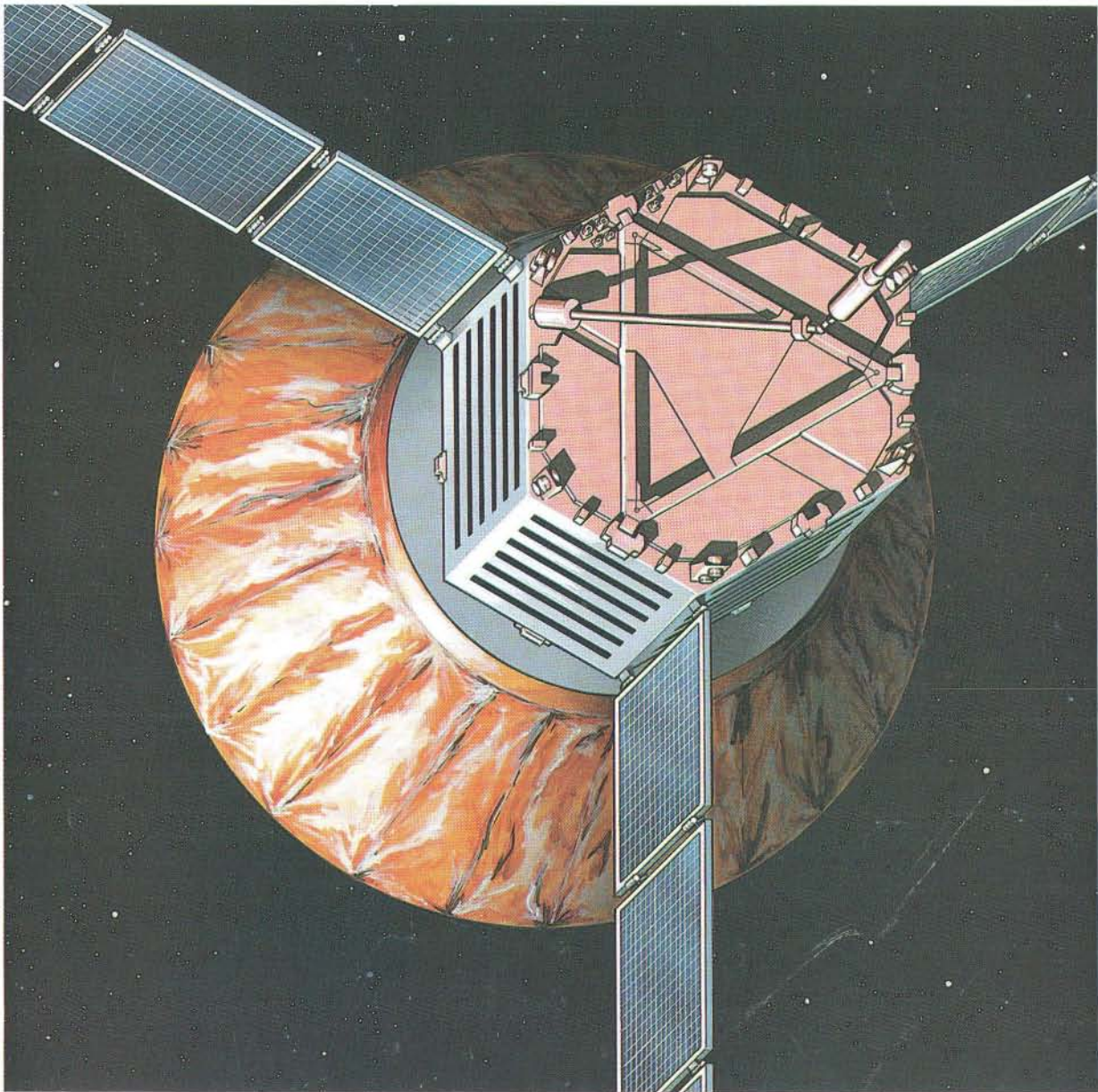
SCIENTIFIC AMERICAN

JANUARY 1990
\$2.95

Can computers think?

Ice ages: a new theory explains the climatic seesaw.

Is the universe right- or left-handed?



*Cosmic Background Explorer will tune in on the big bang
in a search for clues to the origin of the universe.*

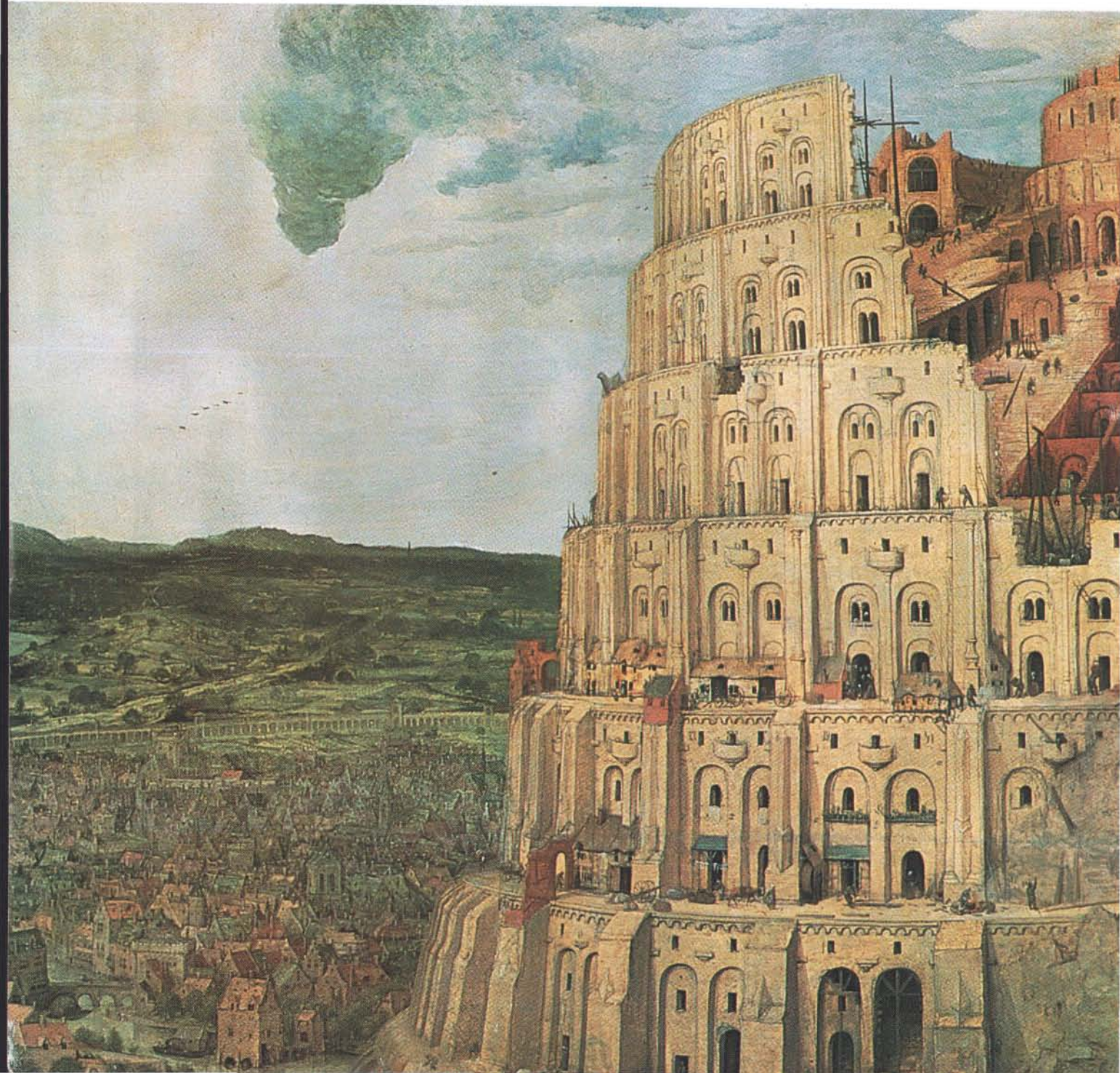
THE TOWER OF BABEL AND SYSTEMS INTEGRATION

The parable is older than writing itself, coming to us from the first murmurings of civilization. Yet its lesson seems to have been aimed specifically at the late twentieth century.

In Genesis, Chapter 11, we read of an unnamed people building a great city on the plain of Shinar (Mesopotamia). To the narrator of this parable, peering across time and desert from his own nomadic traditions, these folk were awesomely clever. They all spoke one common language, and because of this, nothing was impossible to them.

The plan of these ingenious people was to erect a

Pieter Bruegel the Elder, c. 1560, Kunsthistorisches Museum, Vienna



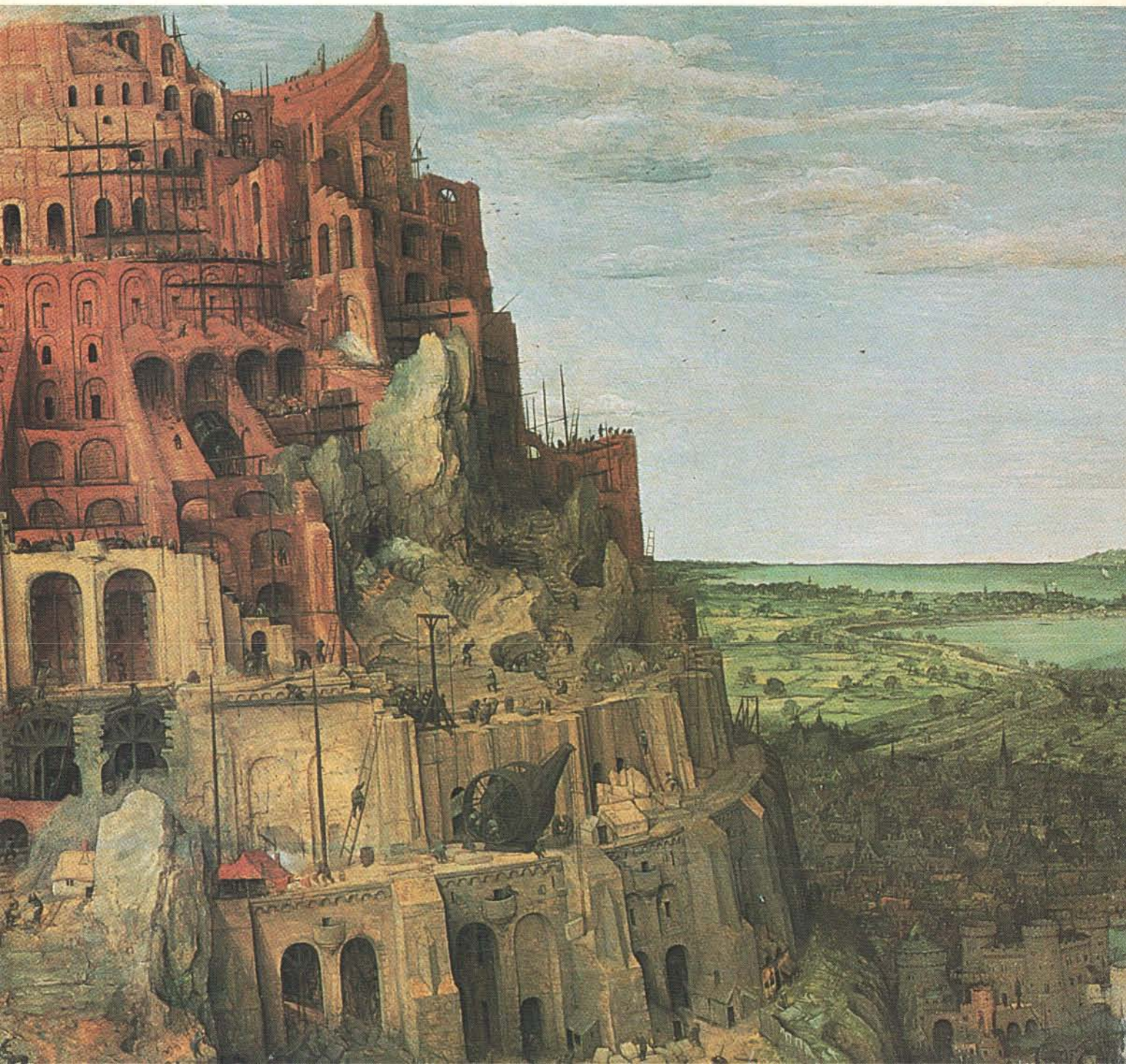
huge temple tower, a ziggurat, whose top would reach into heaven. It was to be an altar to their own intellect and would be called Babel, or "Gate of God." But God himself came down and walked the streets of their city and saw their project under construction. The hubris of this arrogant race angered him. He passed his hand over the city and cursed it. Now where there had been one language were suddenly hundreds. Confusion reigned. Nothing was possible. The people abandoned their city and scattered across the land, taking with them their bewildering tongues. And their vaunted temple, the Tower of Babel, was left untopped; carrion for the wind.

The lesson taught by this ancient parable is uncannily prescient for us in the twentieth century. The revolution in information technology during the past four decades has brought with it the ancient curse of Babel.

Every year witnesses the birth of new computer companies, all fiercely competing with faster, more powerful hardware, new formats and new languages. All contributing to an atmosphere of discord that the narrator of the Biblical story would have had no trouble recognizing, despite the great gulf of time.

Recognizing this discord, Lockheed has a solution; systems integration. For years the company has been synthesizing apparently incompatible systems, whether for use in space, the military, or private industry. To this end, Lockheed has actually been able to work against the Babel effect. And with everyone once again speaking the same language, who knows what wonders are possible?

 **Lockheed**
Giving shape to imagination.



19



ARTIFICIAL INTELLIGENCE: A DEBATE

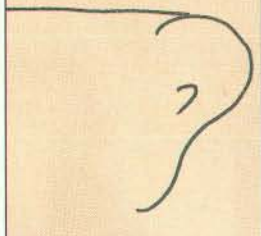
20

Is the Brain's Mind a Computer Program?

John R. Searle

Many people working in artificial intelligence believe that a computer simulation of mental processes could actually think. The author argues that computer programs merely manipulate symbols, without reference to meaning, and so are fundamentally incapable of understanding.

26

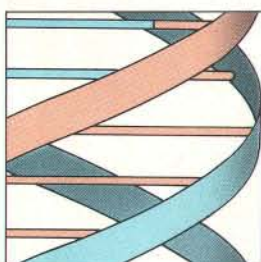


Could a Machine Think?

Paul M. Churchland and Patricia Smith Churchland

Machines that manipulate symbols according to rules may well never achieve intelligence, but, the authors argue, the proposition does not have absolute force. New kinds of systems (such as artificial neural networks) whose physical organization mimics that of the brain might well succeed.

34

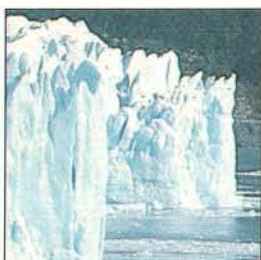


Antisense RNA and DNA

Harold M. Weintraub

A cell translates code-carrying "sense" RNA into protein. Some cells also make "antisense" RNA, which can bind to a particular messenger and thwart translation. In the laboratory, such a molecule can block the expression of a gene and thus reveal the gene's function. In the future, antisense molecules might be recruited to turn off viral genes.

42



What Drives Glacial Cycles?

Wallace S. Broecker and George H. Denton

Astronomical changes are ultimately responsible. Their effect, though, is to alter the intensity of summer sunlight in the northern latitudes. How are the astronomical changes converted into global climatic changes that trigger ice ages? The authors think the variations in the heat of northern summers force a worldwide reorganization of the ocean and atmosphere.

98



The Handedness of the Universe

Roger A. Hegstrom and Dilip K. Kondepudi

From electrons and atoms to molecules, from DNA and proteins to spiraling vines and seashells and on to human beings, nature exhibits handedness, or chirality. The preference for left- or right-handedness seems to be related to fundamental asymmetries in the universe at the atomic scale, but the cause-and-effect relations have yet to be figured out.

106

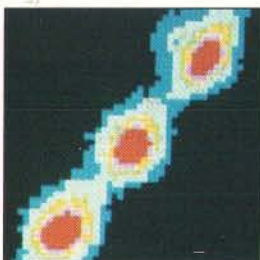


Stress in the Wild

Robert M. Sapolsky

The proper study of humankind may be the baboon, at least with respect to understanding the hormonal effects of stress. Observations of the olive baboon in an African wildlife preserve support the notion that personality strongly influences the hormonal response to stress and, in doing so, influences vulnerability to stress-related disorders.

114



Microplasmas

John J. Bollinger and David J. Wineland

Strip electrons from some thousands of atoms, confine the atoms in an electromagnetic trap and cool them to about absolute zero, and you have a microplasma. It forms strange states of matter—sometimes resembling a solid and sometimes a liquid—that offer physicists a new way to investigate fundamental theories of atomic structure.

122



The Cosmic Background Explorer

Samuel Gulkis, Philip M. Lubin, Stephan S. Meyer and Robert F. Silverberg

The satellite, launched late in 1989, may revolutionize our view of the origin and the fate of the universe. Scanning the skies from an earth orbit high above the obscuring atmosphere, its sensitive instruments will be measuring microwave radiation left over from the big bang and looking for infrared radiation from the very first generation of stars.

DEPARTMENTS

6 Letters

7



50 and 100 Years Ago

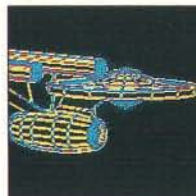
1890: The Forth Bridge in Scotland, just completed, towers 361 feet.

8 Science and the Citizen

94 Science and Business

130 The Amateur Scientist

136



Computer Recreations

Two software packages can generate cellular automaton worlds.

140 Books

144 Essay: *Richard Elliot Benedick*

Will this become the only way to save the rain forests?



Fifty-four irreplaceable acres of tropical rain forest are wiped out every minute. Yet we need every acre desperately.

Rain forests house more than half the world's animals and plants. And they're crucial to maintain the earth's fresh water supply. Without rain forests, we could be living in a burning, barren desert.

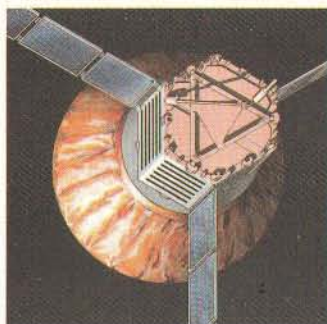
That's why World Wildlife Fund needs your help to save the rain forests and thousands of species that call them home. We're fighting for their survival and, ultimately, our own.

So send for information now. Help us save life on earth. Otherwise, better save this picture.

World Wildlife Fund

Dept. A, 1250 24th St. NW, Washington, D.C. 20037
Prepared as a public service by Ogilvy & Mather.





THE COVER painting depicts NASA's *Cosmic Background Explorer*, which recently began to collect data bearing on the early history of the universe (see "The Cosmic Background Explorer," by Samuel Gulkis, Philip M. Lubin, Stephan S. Meyer and Robert F. Silverberg, page 122). COBE's instruments will scan the sky for a year in an effort to map the microwave background radiation that was emitted soon after the big bang and infrared radiation that is expected to have survived from the earliest stars.

SCIENTIFIC AMERICAN®

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Armand Schwab, Jr., *Managing Editor*; Timothy Appenzeller, Laurie Burnham, *Associate Editors*; Timothy M. Beardsley; Elizabeth Corcoran; John Horgan; June Kinoshita; Philip Morrison, *Book Editor*; Corey S. Powell; John Rennie; Philip E. Ross; Ricki L. Rusting; Russell Ruthen; Paul Wallich; Karen Wright

ART: Samuel L. Howard, *Art Director*; Edward Bell, Joan Starwood, *Associate Art Directors*; Johnny Johnson

COPY: Maria-Christina Keller, *Copy Chief*; Nancy L. Freireich; Michele S. Moise; Philip M. Yam

PRODUCTION: Richard Sasso, *Vice President Production and Distribution*; Managers: Carol Eisler, *Manufacturing and Distribution*; Carol Hansen, *Electronic Composition*; Leo J. Petruzzi, *Manufacturing and Makeup*; Carol Albert; Madelyn Keyes; William Sherman

CIRCULATION: Bob Bruno, *Circulation Director*; Lorraine Terlecki, *Business Manager*; Cary Zel, *Promotion Manager*

ADVERTISING: Robert F. Gregory, *Advertising Director*. OFFICES: NEW YORK: Peter Fisch; John Grant; Meryle Lowenthal; William Lieberman, Inc. CHICAGO: 333 N. Michigan Avenue, Chicago, IL 60601; Patrick Bachler, *Advertising Manager*; Litt Clark, *Midwest Manager*. DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; William F. Moore, *Advertising Manager*; Edward A. Bartley, *Detroit Manager*. WEST COAST: 1650 Veteran Avenue, Suite 101, Los Angeles, CA 90024; Kate Dobson, *Advertising Manager*; Joan Berend, San Francisco. ATLANTA, BOCA RATON: Quenzer/Stites. CANADA: Fenn Company, Inc. DALLAS: Griffith Group.

ADVERTISING SERVICES: Laura Salant, *Sales Services Director*; Diane Greenberg, *Promotion Manager*; Ethel D. Little, *Advertising Coordinator*

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager*, London; GWP, Düsseldorf. HONG KONG/SOUTHEAST ASIA: C. Cheney & Associates. SEOUL: Biscom, Inc. SINGAPORE: Cheney Tan Associates. TOKYO: Nikkei International Ltd.

PUBLISHER: John J. Moeling, Jr.

SCIENTIFIC AMERICAN, INC.

415 Madison Avenue
New York, NY 10017
(212) 754-0550

PRESIDENT AND CHIEF EXECUTIVE OFFICER:
Claus-Gerhard Firschow

EXECUTIVE COMMITTEE: Claus-G. Firschow; *Executive Vice President and Chief Financial Officer*, R. Vincent Barger; *Vice Presidents*: Linda Chaput, Jonathan Piel, Carol Snow

CHAIRMAN OF THE BOARD:
Georg-Dieter von Holtzbrinck

CHAIRMAN EMERITUS: Gerard Piel

THE ILLUSTRATIONS

Cover painting by Hank Iken

Page	Source	Page	Source
19-24	Michael Crawford	108	Patricia J. Wynne, Gabor Kiss (top), Gabor Kiss (bottom)
27	Patricia J. Wynne	109	Gabor Kiss
28-30	Andrew Christie	110-111	Robert M. Sapolsky
35	Harold M. Weintraub, Jonathan G. Izant	112	Gabor Kiss
36-37	Andrew Christie	115	John J. Bollinger, David J. Wineland (left), George V. Kelvin (right)
38	Joseph N. M. Mol, Vrije University, Amsterdam	117	George V. Kelvin
39	Andrew Christie	118	John J. Bollinger, David J. Wineland
42	George H. Denton	119-120	George V. Kelvin
44-47	George Retseck	123	Ian Worpole
48	Bruce Cornet, Lamont-Doherty Geological Observatory of Columbia University (left), Dee L. Breger, Lamont-Doherty (center and right)	124	Philip M. Lubin, University of California, Santa Barbara
49-50	George Retseck	125-129	Ian Worpole
99	Enid Kotschnig	131-134	Michael Goodman
100-105	Hank Iken	136-137	Edward Bell
107	Robert M. Sapolsky	138-139	Denise Saylor

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Copyright © 1989 by Scientific American, Inc. All rights reserved. Printed in the U.S.A. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Subscription rates: one year \$27, two years \$48, three years \$66 (outside U.S. and possessions add \$11 per year for postage). Subscription inquiries: U.S. only 800-333-1199; other 515-247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51593.

LETTERS



To the Editors:

In "The Metamorphosis of Information Management" [SCIENTIFIC AMERICAN, August, 1989], David Gelernter's comparison between software and construction projects is both misleading and illuminating in regard to the current state of software. No one buys a building with the expectation of another "new release" in six months, and yet that practice is common for software.

No engineer would try to construct a bridge—or the Eiffel Tower—by piling together pieces without analyzing their interactions and behavior as a whole. Gelernter advocates building information refineries containing "thousands or even tens of thousands of modules...[in a structure that] no single programmer would understand...in its entirety."

Programs—because of their discrete behavior and large number of states—are more difficult to analyze than bridges. Parallel computers exacerbate this problem by introducing nondeterminism. Better techniques of program specification and verification are necessary if computer programs are to become as reliable as bridges.

JAMES R. LARUS

Computer Sciences Department
University of Wisconsin
Madison

The author responds:

No doubt about it, programs are *not* exactly like bridges, and I thank Professor Larus for pointing this out. But I have to differ with him on the rest of his letter, which raises by implication some very important issues for software and computer science.

I think Professor Larus is implying that if we can't analyze a structure, we shouldn't build it. This is a debatable proposition. My own feeling is that it is profoundly wrong.

Since Larus mentions bridges, let's talk about bridges. In *The Tower and the Bridge* (Princeton University Press, 1985), David P. Billington discusses precisely this issue: Should bridge designers be constrained by the limits of available analytic techniques? On page 151 he writes

that the emphasis placed by the Germans on analysis in the late 19th century "drew them away from forms for which they had no calculations, and thus narrowed the range of structural possibilities." Billington points out (on page 163) that in the U.S. of the 1930's, engineers were "led away from the possibilities for new forms" involving deck-stiffened arches by the complexity of their analytic models. He writes (on page 10), "The fact that these works need not—indeed, in some cases should not—be based on general theories is apparent from concrete studies in the history of technology."

The point is fundamental and transcends bridges, software and whatnot: throughout the history of engineering, imagination has wildly outstripped analysis. Had builders traditionally accepted the limitation that you may build only what you can analyze, our culture would have been immeasurably the worse off for it. The correct analysis of Roman and Byzantine domes or Gothic fan vaulting still occasions technical debate. Historically, engineers have relied on meticulous observation, detailed experiments, a deep knowledge of their materials and, above all, a sense of form—supported (it goes without saying) by the best mathematical tools available, as long as they are of some practical use. We in software strive to do the same.

Analysis is catching up, but in software it still lags far behind. As Larus says, "better techniques of program specification and verification" are indeed necessary. But we refuse categorically to sit on our hands until those techniques appear, nor (to be frank) have we felt limited or constrained in any way by their absence.

Nondeterminism is a case in point. It greatly complicates the formal analysis of program behavior and in theory might make programs harder to write. But in our experience, it more often makes the job easier. Parallel programs pose some special debugging problems, but (again) in our experience, nondeterminism isn't one of them. We are eager to learn as much as we can from the "theoretical systems" people. But we will be limited in our own work strictly by the problems we can't solve, not by the problems they can't solve. I hope and expect that the leading edge of this field will continue to depend only on the imagination of our best builders.

DAVID GELERNTER

Department of Computer Science
Yale University
New Haven, Conn.

To the Editors:

In "The Quiet Path to Technological Preeminence" [SCIENTIFIC AMERICAN, October, 1989], Robert B. Reich writes, "Most U.S. corporate researchers and design engineers work in laboratories that are separated geographically as well as culturally from the factories, warehouses and distribution facilities where their ideas might eventually be implemented. Research facilities typically occupy modern, campuslike buildings in bucolic surroundings.... R&D often has relatively little connection to the rest of a company's undertakings."

Your magazine chose to illustrate Mr. Reich's point with a photograph whose caption asserts: "Pastoral setting of many U.S. research centers (such as this AT&T Bell Laboratories facility in Holmdel, N.J.) fosters creativity, but the separation of research laboratories from the factory floor can make U.S. companies less efficient than their Japanese counterparts at translating new discoveries into profitable products and the processes for producing them."

You selected the wrong company to make the point.

For decades Bell Labs has pioneered information technology by "co-locating" its R&D people at manufacturing sites. More than 3,000 R&D scientists and engineers work hand in hand with manufacturing engineers and marketing personnel at AT&T factories in six states to ensure the timely, cost-effective introduction of AT&T products and services.

At Holmdel, N.J., where some of our fundamental research is performed, Bell Labs researchers and developers collaborate daily with co-located AT&T product and marketing employees on products ranging from residential phones to new computer architectures.

Reich is right in saying a gulf will develop between R&D on the one hand and manufacturing and marketing on the other unless corporations work assiduously to bridge those functions. At AT&T, that's what we have done.

S. J. BUCHSBAUM

Executive Vice President
AT&T Bell Laboratories
Holmdel, N.J.

Every month we receive hundreds of letters from our readers. We and our authors thank you for sharing your thoughts with us—and ask for your forbearance. The sheer number of letters makes it impossible for us to answer more than a fraction of them.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JANUARY, 1940: "Down the lower side of a cigar-shaped radio beam, an airplane flown by Capt. Milton M. Murphy and Jack Haynes, Civil Aeronautics Authority inspector, glided repeatedly to safe landings on East Boston airport in Massachusetts recently. These flights demonstrated for the first time an application of the klystron, an invention developed at Stanford University, which generates ultra-short radio waves that can be directed as a searchlight directs a beam of light."

"When the government-sponsored South Pole Expedition commanded by Admiral Byrd arrives in the Antarctic sometime this winter, it will have a unique vehicle, which will be used as a mobile base. Designed by Dr. Thomas C. Poulter, it looks like a cross between a huge bus and a military tank. It is so constructed that it can negotiate rough ice fields, cross crevasses up to 15 feet wide, make speeds up to 25 miles an hour, and have a cruising range of between 5000 and 6000 miles. The Snow Cruiser will be 55 feet long and 15 feet wide, and will carry on its roof a five-passenger airplane."

"Experimental use of the parachute is being made to fight forest fires in the Chelan National Forest in Washington State and, while definite conclusions have not yet been drawn, the experiments are highly promising. Firefighters will be intensively trained and will be dropped on the 'softest' spot as near the location of the fire as possible. Upon landing, the man crawls out of his protective suit, frees himself of the parachute, and strikes out for the nearby fire."

"Germs of two varieties of *Streptococcus salivarius* were found on the rims of glasses in taverns, restaurants, cocktail lounges, and soda fountains. The germs, harmless in themselves, can be used as an index of how well such glasses are washed between drinks and of how many more dangerous germs may be left on the glasses by careless washing. Of all the glass-

es examined, those in soda fountains were the least germ."

"During daylight hours the dirigibles of the Goodyear Tire & Rubber Company provide a unique way of seeing the sights of New York City from the sky, but at night they carry over the city an ingenious neon advertising sign."



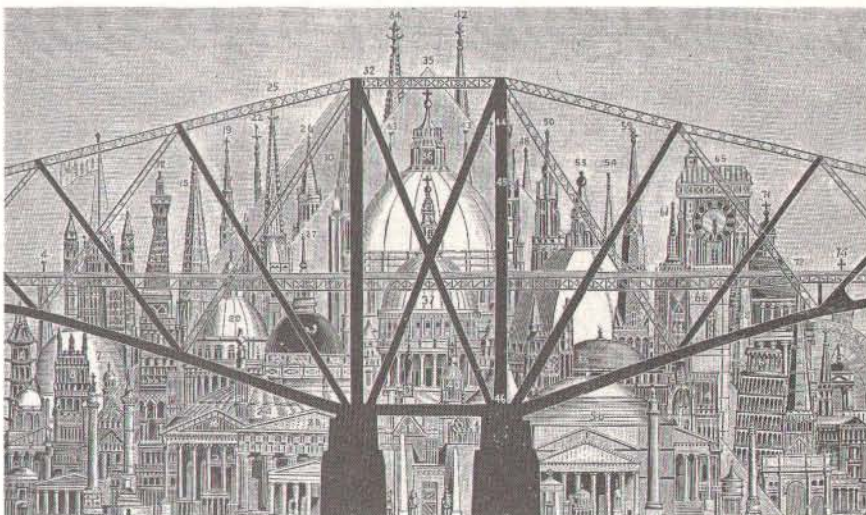
JANUARY, 1890: "Miss Helen A. Keller, who is at present an inmate of the Perkins Institute for the Blind in Boston, was deprived of her sight and hearing at the age of eighteen months. At the age of six, being deaf, dumb, and blind, she was put under the charge of Miss Annie M. Sullivan, who undertook to instruct her in the touch alphabet, and so eager was her pupil for knowledge, and so quick of perception, that she now is able to read and write with perfect facility. It will be a matter of the profoundest interest to watch the development of human nature uninfluenced by the usual surroundings of life, and to watch the soul expand and grow by its own virility."

"The use of the Colorado Midland's rotary snow shovel on the Denver, Texas, and Fort Worth seems to have created a mild sensation. A local paper says: 'It was put to work in a big cut where the snow was about 20 feet deep. Around the center of the cut, a strange sight was witnessed. Those who were standing on either side of the plow were suddenly deluged with

a shower of beef steaks. On all sides fell porterhouse, sirloin, round steaks, small steaks, shoulder steaks, with occasionally a slice of liver or a nicely cut rib roast. Investigation disclosed the fact that a herd of Texas cattle had crowded into the cut and had frozen and been buried in the drifts."

"After showing that friction makes perpetual motion impossible, Professor Hele Shaw reflects upon the state of affairs that would follow if friction were to cease to act. The whole force of nature would be at once changed, and much of the dry land and most of our buildings would disappear beneath the sea. Such inhabitants as remained a short time alive would not only be unable to provide themselves with fire or warmth, but would find their very clothes falling back to the original fiber from which they were made; they would also be unable to obtain food, from inability to move themselves by any ordinary method of locomotion, or, what would be equally serious, having once started into motion, from being unable to stop except when they came into collision with other unhappy beings or moving bodies."

"The Forth Bridge has lately been completed and is now receiving the finishing touches. The bridge is the most important link in the direct railway communication between Edinburgh on the one hand and Perth and Dundee on the other. The total length of the viaduct (the longest in the world) is 8,296 feet, or nearly 1⁵/₈ miles; the headway for navigation is 150 feet; and the extreme height of the structure is 361 feet."



Heights of great buildings of the world and the Forth Bridge compared

SCIENCE AND THE CITIZEN

Aftershocks

California quake intensifies pressure for prediction

The biggest news about the earthquake that shocked California last October may be that it was hardly a surprise—at least not to seismologists. In 1988 the U.S. Geological Survey had flagged virtually the exact segment of the San Andreas fault that failed. It runs near Loma Prieta Mountain in the southern Santa Cruz range. There was, USGS workers said, a 30 percent chance that an earthquake of 6.5 magnitude would happen there within 30 years—one of the highest probabilities in the whole San Andreas fault system.

The Loma Prieta earthquake represents a "vindication of the whole approach" of studying past seismicity to make long-term forecasts of earthquake likelihood, maintains Allan G. Lindh of the USGS in Menlo Park, Calif. Moreover, in the months before the quake there were clues that a stretch of the San Andreas that had been quiescent since 1914 was preparing to move again, suggesting that some nearer-term forecasting might be possible. "It is becoming a science," says James H. Dieterich of the USGS.

Current theory suggests big earthquakes occur where stress caused by tectonic plates grinding against one another accumulates, producing a "seismic gap," instead of being relieved by many small earthquakes. Lindh expressed apprehension about the long-silent southern Santa Cruz segment last summer, after an earthquake in August near Lake Elsinore relieved only a little stress.

In retrospect, it seems that that earthquake, together with one that occurred nearby in June, 1988, "in some sense foretold" the Loma Prieta event, Lindh says; "it's hard for most people to believe there wasn't some physical process occurring at depth... starting in June of 1988." Distance measurements made with lasers atop Loma Prieta indicate there were indeed "marginally significant" changes in the rate of ground movement there starting in June, 1988, according to USGS's James C. Savage.

What would be needed to try to improve near- or intermediate-term forecasting? Better seismometers would help reveal how earthquakes propa-



11 PHYSICAL SCIENCES 13 BIOLOGICAL SCIENCES 15 MEDICINE 16 PROFILE

gate. There are hundreds of seismometers in California, but most of them are inferior to the best modern instruments. "Work is limited by the lack of good data," says William L. Ellsworth of the USGS.

USGS researchers think the best hope for monitoring any small telltale ground movements that might precede an earthquake would be a combination of more borehole strain gauges and better use of the Global Positioning System. This system, which utilizes military satellites, can measure horizontal ground movements of less than one centimeter over a distance of 40 kilometers. Measurements are made infrequently, however, and coverage is sparse. Receivers cost up to \$80,000, and many more would be needed to achieve continual monitoring of active faults near populated regions of California.

Would the public react sensibly to shorter-term earthquake forecasts, or is mass panic likely? No one can be sure, but Lindh is encouraged by the "intelligent and responsible" actions of state officials, who, after consulting with the USGS and the California Earthquake Prediction Evaluation Council, have in recent years been issuing short-term advisories of increased earthquake likelihood after small earthquakes (including the Lake Elsinore event). According to Richard A. Andrews of the California Office of Emergency Services, some useful preparatory actions have been taken as a result of the advisories: emergency generators were checked, people were reminded where to find the American Red Cross and some fire trucks were rolled out. These advisories "marked a sort of watershed," Lindh says. "People could not conceive of doing that 10 years ago."

Investment in improved warning

systems might be timely. Loma Prieta "filled in" one big seismic gap on the San Andreas; the next conspicuous gap, according to Lindh, lies farther north, only some 20 to 40 kilometers south of San Francisco. Most USGS workers feel that Loma Prieta shifted stress on the San Andreas farther north, making a full-size successor to the Little Big One more likely.

The rapid, worldwide growth of cities near seismically active zones makes the improvement of predictive capability pressing, argues Roger Bilham of the University of Colorado at Boulder. Urgent as it is, it may not be accomplished soon. According to Robert L. Wesson, director of the USGS's Office of Earthquakes, Volcanoes and Engineering, the USGS's funding for earthquake-related research has fallen by 40 or 50 percent in real terms since 1978.

—Tim Beardsley

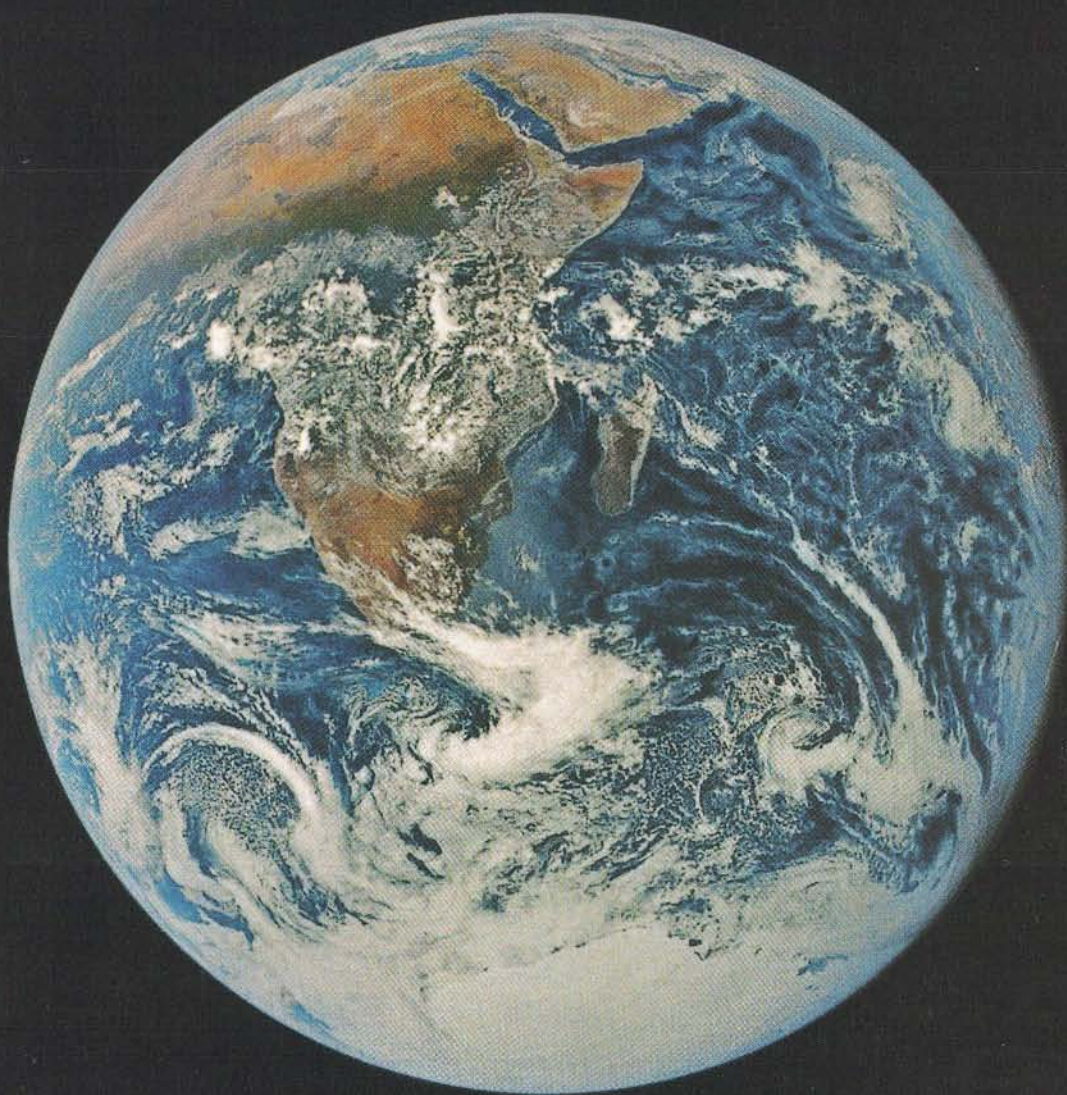
Loud and Clear

Science's new voice at the White House speaks out

Allen Bromley, director of the White House's Office of Science and Technology Policy and Assistant to the President for Science and Technology, sounded a clear warning in November that budget pressures are likely to mean delaying some of the "big science" projects initiated during Ronald Reagan's terms of office.

Speaking in forthright fashion at a meeting of the D.C. Science Writers' Association, Bromley singled out such megaprojects as the Superconducting Supercollider, the National Aerospace Plane, the Human Genome Project and the space station. Unless watched carefully, he said, they could displace small science carried out by individual investigators, and he is "committed to making sure that doesn't happen." Expanding on the topic, he said that the big science projects could not all be continued in parallel at the level sought by their proponents: "If indeed we do them all, we can't do them all at the same time...we don't have the funds."

Bromley also advocated increasing the proportion of the Defense Department's research budget devoted to basic research, which is currently some 8 percent. He said he would argue that



**IF YOU'RE NOT RECYCLING
YOU'RE THROWING IT ALL AWAY.SM**

A little reminder from the Environmental Defense Fund that if you're not recycling, you're throwing away a lot more than just your trash.

You and your community can recycle. Please write the

Environmental Defense Fund at: EDF-Recycling, 257 Park Avenue South, New York, NY 10010, for a free brochure that will tell you virtually everything you need to know about recycling.



in the current atmosphere of reduced international tension, investment in basic research is necessary to keep the military from being technologically blindsided.

In response to a question, Bromley said he had not been consulted by Secretary of Health and Human Services Louis W. Sullivan about the decision to make permanent a ban on federal support for research that involves transplanting fetal tissue from induced abortions into human recipients. Bromley said he was "concerned" that views on abortion "could become a litmus test" for government appointments in science and technology but added that he is not aware that such a test has been applied.

Bromley rejected the suggestion that the administration is dragging its feet over an international agreement limiting emissions of carbon dioxide. Decisions should await the assessments of greenhouse warming being carried out under the auspices of the United Nations, he said. "By fall 1990 we will be in a position to discuss how we could stabilize or even reduce greenhouse-gas emissions." —T.M.B.

Nit Picker

Ancient combs served more than cosmetic purposes

Although St. Francis of Assisi once characterized lice as "pearls of poverty," the ancient dwellers of the Mediterranean took a more worldly view of *Pediculus capitis*. Egyptian priests, resorting to a scorched-earth policy, simply shaved their heads and bodies to prevent infestation. By Roman times a more subtle defense—one that is still in use—was developed: the nit comb.

Writing in *Biblical Archaeology Review*, Kostas Y. Mumcuoglu, a parasitologist at the Hebrew University of Jerusalem, and Joseph Zias, a curator at the Israel Department of Antiquities and Museums, report that they found lice of every developmental stage from egg to adult in organic material taken from two wooden combs found at Qumran, near the site where the Dead Sea Scrolls were discovered. That area had been populated by the Essenes, a Jewish sect, until it was laid waste by the Roman army (A.D. 68) during the First Jewish Revolt. The authors also found eight lousy combs among artifacts that had been taken from sites dating to the second revolt (A.D. 132-135).

Mumcuoglu and Zias maintain that



WOODEN COMB from the second century was preserved in the Judean desert. Head louse was taken from a first-century comb found at Qumran.

the combs were intended as delousers: a coarse-toothed edge served to straighten the hair; a fine-toothed edge served to delouse it. The design appears to have been effective, as many lice found on the combs had been broken into pieces.

These organic specimens could yield information on matters beyond the history of hygiene. Zias says he is looking for someone sufficiently versed in microsurgery to open the gut of one of the ancient lice and extract any human blood it may contain. He says the nucleus from even a single leukocyte (white blood cell) might furnish geneticists with enough DNA to compare its ancient donor with modern populations. Such comparisons might gauge the genetic link between ancient and modern Jews and perhaps suggest ancestral ties with other peoples around the world. Such findings would supplement the chronological study of human genetics

that began with experiments conducted on Egyptian and South American mummies.

"Or we could try to clone the human louse itself," Zias says. "Morphologically it has not changed, but who knows about its molecular biology." Indeed, with its generations measured in weeks rather than decades, the lowly louse will have experienced far more evolutionary time than humans have over the past 2,000 years.

—Philip E. Ross

Nanofuture

How much fun would it be to live forever?

K Eric Drexler is a prophet who likes to think small. His vision is of a world remade by hordes of self-replicating microscopic robots. Drexler's search for "robust arguments for what is possible" has led him to proclaim (at fees that reach \$5,000 for an appearance) that nanometer-scale devices will someday give human beings a revolutionary power: inexpensive control over the structure of matter. In the nanotechnology age most international trade will be unnecessary, since anything that can be made will be "grown" from feedstock by nanomachines programmed by onboard computers. In Drexler's world people will live indefinitely, their illnesses corrected in situ by nanorobots executing molecular adjustments. Yet life in the nanofuture will not be all roses, Drexler fears: "gray goo" weapons might be developed—soups of omnivorous nanorobots that could reduce the biosphere to dust.

Far out? No farther out than Stanford University, where Drexler is a visiting scholar in the computer science department. His Foresight Institute, together with the Global Business Network (an eclectic group comprising, among others, Peter Schwartz, a former Shell executive, the biologist Lynn A. Margulis and rock musicians Brian Eno and Peter Gabriel), recently held an international conference in Palo Alto, Calif. Its purpose: to enable experimental scientists, venture capitalists and armchair nanoenthusiasts to start planning the nanotechnology revolution.

Drexler looks chiefly to biochemical engineering for the molecular components that will constitute the first assembler: a programmable nanomachine tool that can build structures atom by atom, including other assemblers. Once the first assembler has

been built, Drexler says, its offspring will be able to create armies of assemblers in days.

Engineering projection, or science fiction? The scientific presentations at the conference were real enough. Tracy Handler of Du Pont described how she and her colleagues had designed, apparently successfully, a protein consisting of four helices that fit together to form a cylinder; they made the protein by inserting a synthetic gene into the bacterium *Escherichia coli*. Jay Ponder of Yale University said he is starting to predict the three-dimensional structures of novel proteins—a formidable problem—by working from analogy with known structures. John Foster of the IBM Almaden Research Center in San Jose told how his group is starting to manipulate individual molecules with a variant of the scanning tunneling microscope.

Others pondered questions that have crossed few human minds. Ralph C. Merkle of the Xerox Palo Alto Research Center (who, as a co-inventor of public-key cryptography, apparently has armchair credentials) proposed that nanomachines be forbidden to have sex, thus preventing them from reshuffling their programs and surprising their creators.

There were a few open doubters. Lester Millbrath of the State University of New York at Buffalo worried that the development of nanotechnology ("the most important moral decision ever made by our species") was unlikely to solve the environmental crisis—a view hotly disputed by nanoenthusiasts. Federico Capasso of AT&T Bell Laboratories questioned the potential of one nanoscale technology, quantum-based microelectronics, by pointing out that only a few percent of discoveries lead to new devices.

Despite Drexler's guess that an assembler might be built during the first third of the next century, some participants said privately that they thought the necessary technologies lie still in the distant future. "He's definitely being overoptimistic," said Michael D. Ward of Du Pont, who designs multimolecular lattices from a knowledge of the component molecules.

Some experts on the very small who were not at the conference also have reservations. Harold Craighead, director of the National Nanofabrication Facility at Cornell University, is enthusiastic about applications in electronics and optoelectronics but is dubious about Drexler's nanorobots. "It takes in general a complicated factory to make a complicated mechanical object, and it gets harder as the object

gets smaller," he says. Richard S. Muller, co-director of the Sensor and Actuator Center at the University of California at Berkeley, which last year fabricated a motor .1 millimeter in diameter, adds that "they haven't come up with anything that's working yet—the substance isn't really there." Rolf W. Landauer of IBM has cautioned that atomic-scale devices in general are too sensitive, too variable and too delicate to carry out computation.

Drexler welcomes such criticism. He argues that the consequences of nanotechnology will be so great that humankind had better start planning now, to maximize the time for debate. "Everything I heard was a new opportunity or a new strength... rather than a new problem," he declares. —T.M.B.

PHYSICAL SCIENCES

Quasicrystal Clear

Is entropy the driving force behind this odd form of matter?

What exactly are quasicrystals? Workers at the National Bureau of Standards discovered this curious state of matter in an alloy of aluminum and manganese in 1984. The alloy's molecular structure displayed fivefold symmetry: it diffracted X rays into patterns that looked the same when rotated by fifths of a circle. According to classical crystallography, one can no more build a crystal out of fivefold symmetric "unit cells" than one can cover a plane with pentagons; in each case, gaps are unavoidable. Physicists agreed to call the new type of matter quasicrystals but disagreed about its nature.

Several researchers, notably Paul J. Steinhardt of the University of Pennsylvania, have maintained from the start that quasicrystals represent a real-world embodiment of a mathematical construct called Penrose tiling. In the early 1970's Roger Penrose of the University of Oxford found that two types of rhombus (a parallelogram whose sides are all equal), if fitted together according to certain "matching rules," can cover an infinite plane while never settling into a repeating, single-celled pattern. Although the pattern formed by the tiles is technically nonperiodic, it is dense with decagons and five-pointed stars—shapes that exhibit fivefold symmetry.

Many solid-state physicists bridled at the Penrose tile model. They said it

was difficult to imagine how Penrose's complicated matching rules would be enforced in nature. They pointed out that the proper placement of a Penrose tile often requires knowledge of the positions of very distant tiles. Nature might accomplish such a trick through a nonlocal effect—in which conditions in one region instantaneously influence events elsewhere. Indeed, Penrose himself has suggested that some nonlocal effect related to quantum phenomena might give rise to quasicrystals. Still, the vast majority of solid-state physicists view nonlocality as unacceptable, more fiction than science.

Some theorists have tried to invent completely local matching rules, which might obviate the need for nonlocal knowledge in Penrose tiling and, perhaps, nonlocal effects in quasicrystals. One such set of rules, advanced by Steinhardt, David P. DiVincenzo of IBM and others, requires that the sides of tiles vary in their "stickiness," or ability to adhere to neighboring facets. Skeptics assert that such matching rules are even more complicated—and so even more unlikely to occur in nature—than Penrose's.

Until recently many physicists preferred to believe a far more conventional alternative to the Penrose tile model: the icosahedral glass model. Advanced chiefly by Alan I. Goldman of Iowa State University and Peter W. Stephens of the State University of New York at Stony Brook, it held that quasicrystals have only occasional islands of structure in a sea of disorder—or, more specifically, icosahedral clusters of atoms scattered at random in a glassy matrix.

Although not as exciting as the Penrose tile model, the icosahedral glass model better explained the blurry X-ray diffraction lines generated by early samples of quasicrystals. (Icosahedrons, which have 20 triangular faces, exhibit fivefold symmetry.) Then numerous groups, the first at Tohoku University in Japan, began producing quasicrystals that displayed far more order than previous samples. An analysis of the new samples last year by Goldman and Stephens, among others, finally ruled out their own icosahedral glass theory. Their work, they say, also eliminates an earlier proposal by Linus Pauling that quasicrystals are just conventional crystals that intersect in unconventional ways.

Meanwhile the stock of another theory has risen. The theory is based on an examination by Michael Widom of Carnegie-Mellon University and Katherine J. Strandburg of Argonne Nation-

al Laboratory of what happens if one assembles Penrose tiles at random, eliminating all rules except the requirement that the tiles cover the plane. The investigators found that in most cases the tiles form a pattern that exhibits the same basic properties of fivefold symmetry and nonperiodicity as perfect Penrose tiles, even though the position of many individual tiles is changed.

The chief advantage of this theory is that it has the powerful backing of the second law of thermodynamics, which holds that nature favors systems possessing greater entropy, or potential randomness. In a sense, entropy, rather than matching rules or nonlocal tricks, becomes the system's governing principle.

Recent experiments by Peter A. Bancel of IBM seem to support such a model. Bancel found that the X-ray diffraction lines of quasicrystals made of aluminum, copper and iron became sharper as the temperature of the materials increased. This counterintuitive phenomenon, Bancel says, is a predicted symptom of a high-entropy system rather than one restricted by Penrose matching rules.

Does that mean that true Penrose quasicrystals—conforming precisely to his matching rules—do not exist? David R. Nelson of Harvard University thinks so. He finds the entropy model so convincing that he considers the search for perfect Penrose quasicrystals a waste of time. Not surprisingly, other investigators disagree. "Finding perfect Penrose tiles is still the Holy Grail," DiVincenzo remarks.

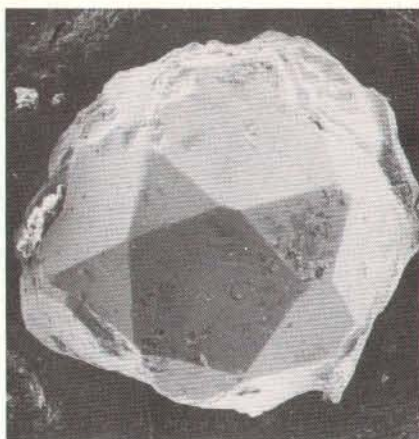
—John Horgan

Bottled Antimatter

New trap cools antiprotons to test nature's antisymmetry

A device tested at CERN, the European laboratory for particle physics near Geneva, may give investigators a glimpse of what an antimatter world might look like. The device cools antimatter to a temperature of a few degrees above absolute zero and stores it for several days at a time. It has already made possible precise measurements of the antiproton's mass and may yield the first antimatter atoms on earth.

The CERN physicists create the raw ingredients by accelerating particles of ordinary matter to high energies and then steering them into a metallic target. The collisions produce particle-antiparticle pairs. The antiparticles



QUASICRYSTAL of aluminum, copper and iron was made by Peter A. Bancel of IBM. It is .3 millimeter wide.

are then siphoned off by magnets and kept in storage rings. Storage rings generate electromagnetic fields that suspend antimatter within a vacuum and prevent it from annihilating matter. The electromagnetic fields also propel the antimatter at nearly the speed of light. The fact that the antimatter is whizzing by at great velocity makes the task of measuring its properties very difficult.

The new storage device—developed by investigators from Harvard University, the University of Mainz in West Germany and the University of Washington—overcomes this obstacle. As reported in *Physical Review Letters*, the device stores antiprotons (the antimatter counterparts of protons) at energies more than 60 million times lower than those in any conventional storage ring.

To begin the new storage process, the workers divert a group of antiprotons from the storage ring. The antiprotons, which are traveling at an average of one tenth of the speed of light (30 million meters per second), enter a sealed, evacuated chamber and crash through an aluminum plate. The collision slows them down to 700,000 meters per second. The chamber also contains a series of hollow, cylindrical electrodes through which the antiprotons pass. The last electrode generates an electric field that turns the antiprotons around.

A hundred billionths of a second later, a voltage is applied to the other cylinders to generate a new electric field. The combination of this electric field and a magnetic field (produced by an exterior superconducting magnet) creates forces that trap the antiprotons. This electromagnetic bottle is a type of Penning trap [see "Micro-

plasmas," by John J. Bollinger and David J. Wineland; page 114].

The captured antiprotons (as many as 60,000 of them) oscillate rapidly in the trap. To cool the antiprotons further, the workers fill the electromagnetic trap with many cold electrons. As the antiprotons collide with the electrons, they cool as the electrons heat. According to one of the principal investigators, Gerald Gabrielse of Harvard, thousands of antiprotons have been cooled to a temperature below 100 kelvins. At that temperature the antiprotons are moving at 1,300 meters per second.

The new technique has enabled researchers to improve measurements of the antiproton's mass by a factor of 50, Gabrielse says. They found that the proton and antiproton have the same mass within about one part per million. This measurement confirms the predicted symmetry between matter and antimatter.

The group plans to isolate a single antiproton in order to measure its mass 1,000 times more accurately. The workers are also attempting to produce and study cold antihydrogen: the combination of an antiproton and an antielectron.

—Russell Ruthen

The Redshift Blues

New redshift theory challenges both physicists and cosmologists

Redshifts are to astronomy what tape measures are to carpentry: a well-understood tool whose validity hardly seems open to question. Modern cosmology is rooted in the belief that essentially all observed redshifts are caused by the Doppler effect, whereby light emitted by a receding object is "stretched out" and so shifted toward the red end of the spectrum. Emil Wolf of the University of Rochester, writing in *Physical Review Letters*, now questions this traditional interpretation by proposing a previously unrecognized mechanism that, he claims, can completely mimic the Doppler effect.

Wolf suggests that light can be redshifted if it passes through a scattering medium in which the index of refraction varies randomly over both space and time. If suitable correlations exist within the medium, the light will change frequencies even though the overall source is at rest. His calculations show that the result can in principle be a Doppler-like shift across the entire spectrum.

Wolf's ideas have generated con-

siderable interest and controversy among his colleagues, in part because the big-bang theory rests on observations of redshifts that are interpreted as evidence that the universe is expanding. Wolf hesitates to suggest that the big-bang concept might be wrong. "It is something that should be looked at," he says cautiously.

More plausible, Wolf believes, is the possibility that his mechanism may explain long-disputed quasar-galaxy pairings. In some instances, quasars and galaxies that appear to be physically associated have vastly different redshifts, which indicates, by conventional reckoning, that they are billions of light-years apart. Many astronomers—including Christopher L. Carilli of Harvard University, who with two colleagues found the most recent such association—dismiss these associations as chance line-of-sight pairings. A few workers, among them Jack W. Sulentic of the University of Alabama at Tuscaloosa, maintain that they reveal either anomalies unexplained by the big-bang theory or else the existence of some kind of non-Dopplerian redshift.

Naturally, Wolf favors the latter interpretation. Matter surrounding quasars is probably anisotropic (not the

same in all directions) because of turbulence or the powerful jets of matter that quasars often emit. Such anisotropy could produce the sort of correlation mechanism that Wolf has studied, generating redshifts that have hitherto been attributed to the quasars' great distance from the earth. If confirmed, this finding could solve another quasar mystery: some quasar jets appear to be expanding faster than the speed of light. If quasars were significantly less distant than generally believed, the jets would be much smaller and their rate of expansion within the cosmic speed limit.

Wolf is searching for more down-to-earth applications, such as methods for correcting satellite-tracking signals and for improving reference standards. He also reports interest from the Department of Defense, which is intrigued by the possibility that artificially induced spectral modulation could provide the ultimate in coded communications. —Corey S. Powell

BIOLOGICAL SCIENCES

On to the Past *A famed "living fossil" may finally face extinction*

Time may finally be running out for the coelacanth, a group of ichthyologists warns. Although the ancient fish is inedible and lives only at depths greater than 70 meters, the investigators—banded together as the Coelacanth Conservation Council—say the high prices that specimens now command as well as improved angling techniques are inducing fishermen in the Comoro Islands in the Indian Ocean to hunt them deliberately.

Coelacanths, which can reach a length of five feet, are believed to be closely related to the evolutionary ancestors of all land animals. They were thought to have gone extinct some 70 million years ago; then, in 1938, a fisherman caught one. The living fossils have been found only in the waters around the Comoros; the number surviving is not known, but according to one of the council's founding members, Michael N. Bruton of the J.L.B. Smith Institute of Ichthyology in Grahamstown, South Africa, the species may number only in the hundreds.

Concern for this fragile population has recently spurred the council to condemn an ambitious plan by the Toba Aquarium in Japan to capture

two live coelacanths, an event the council fears could trigger global competition. Hajime Nakamura, director of the coelacanth project at the aquarium, says that if coelacanths can be caught and kept alive, two specimens will be brought back to Japan to be studied in preparation for a captive-breeding program. Council members question the feasibility of the plan, pointing out that coelacanths can survive for only a few hours after being brought to the surface (probably because the fish are sensitive to temperature and pressure changes). Nakamura replies that his institution has developed a trap that provides a life-support system. He will not reveal how it works.

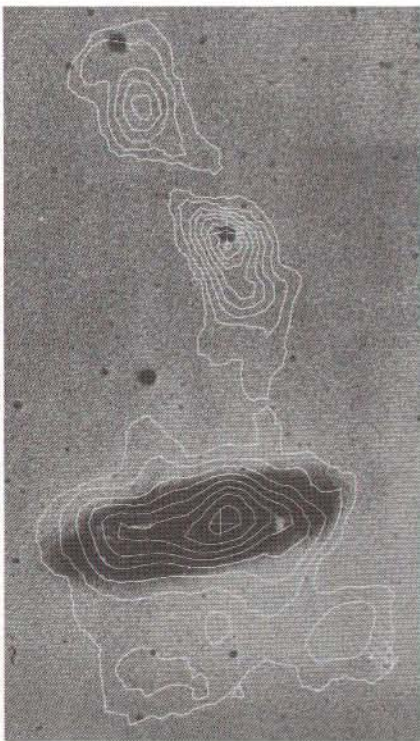
The council doubts Nakamura's assurance. "If they are truly involving themselves in a scientific exercise the aquarium shouldn't be withholding information," Bruton argues. Bruton and colleagues remain unpersuaded on another count as well: because the coelacanth takes from 10 to 15 years to reach maturity, "we think captive propagation is absolute nonsense," he says, adding that "the coelacanth's status is so critical that we cannot afford to risk killing [coelacanths] to transport them halfway round the world." The ichthyologists favor intensive study and conservation of the known wild populations instead.

The coelacanth's defenders note that it was recently listed in Appendix I—the most protected category—of the Convention on International Trade in Endangered Species. Yet Nakamura disputes the coelacanth's categorization as a rare species and says the CITES listing will not affect his project. The council's denunciation is "nothing but harassment resulting from the consciousness of territory of scholars," he charges. —T.M.B.

A Grave Tale

Do whale remains help life spread on the deep-sea floor?

Just over a decade ago most scientists thought the deep-sea floor was a cold, dark desert. Then, in the late 1970's, researchers found that underwater hot springs called hydrothermal vents support oases of life thousands of meters below the ocean's surface. These ecosystems rely not on sunlight for their primary sustenance but on chemicals spewed out by the vents; the chemicals nourish bacteria, which in turn anchor a food chain that includes unique spe-



RADIO-EMITTING CLOUDS of hydrogen seem to connect a quasar and galaxy (crosses mark the objects' centers), although redshifts indicate the objects are nine billion light-years apart.

cies of clams, worms and other fauna.

More recently similar chemosynthetic communities have been discovered around so-called cold seeps, where petroleum and other energy-rich substances leak from below the ocean floor. Although many vents and seeps have now been identified, most cluster within relatively narrow zones—usually midocean ridges or the margins of continents. How did the chemosynthetic creatures disperse to these widely separated zones? A serendipitous encounter by the research submarine *Alvin* has identified a possible stepping-stone—the graves of whales.

The encounter occurred two years ago as researchers were making the last of a series of dives in the Santa Catalina Basin, some 35 kilometers southwest of Los Angeles. Taking photographs at a depth of 1,240 meters, they noticed what looked like an outcrop of limestone. On closer inspection, the formation turned out to be the skeleton of a blue or fin whale 20 meters long. "It was purely a lucky find," says Craig R. Smith of the University of Hawaii at Manoa.

Smith and other investigators from Hawaii and the University of Wash-

ington returned to the site several times to take measurements and collect samples. The skeleton is surrounded by thick mats of bacteria and by clams and mussels similar to those that surround vents and seeps. No flesh is visible, but Smith suggests that the skeleton alone contains enough oil, lipids and other organic materials to support the ecosystem now and possibly for years to come. "There is a lot of energy in the bones," he says.

How old is the carcass? A spike of radioactive carbon 14 in the bones indicates that the whale died sometime after the superpowers began large-scale atmospheric testing of nuclear weapons; that means the carcass must be less than 34 years old, according to Smith. The size of the clams on the carcass provides a lower age limit of at least three years.

In a letter to *Nature*, Smith and his colleagues suggest that a typical whale carcass could support a chemosynthetic ecosystem for at least five years. They estimate that deaths among gray whales alone could create at least 500 new deep-sea habitats a year in the North Pacific. Before the advent of whaling, of course,

that number would have been much greater. —J.H.

See Spot See Blue

Curb that dogma!

Canines are not colorblind

Perhaps because we envy their superior olfactory and auditory talents, we humans have long denigrated the eyesight of dogs. We have assumed that they inhabit a drab, black-and-white world, devoid of color. Now experiments done at the University of California at Santa Barbara debunk this myth: although our color vision is more acute than theirs, our faithful friends can easily tell a blue ball from a red one.

In the experiments, reported by Jay Neitz, Timothy Geist and Gerald H. Jacobs in *Visual Neuroscience*, dogs viewed three screens illuminated from behind by colored lights. Two of the screens displayed the same hue and the third a different one. The scientists trained the dogs to choose the odd-colored screen by poking it with their snouts. If they made the right choice, they received a cheese-and-beef-flavored pellet.

The three dogs tested—two greyhounds and a toy poodle—easily discerned the difference between white and colored light and between colors at opposite ends of the spectrum, and they rivaled humans in telling apart closely related hues of violet and blue. Yet the dogs could not discriminate among colors from greenish-yellow through orange to red.

The investigators concluded that dogs have two types of photoreceptors: one responds only to the blue-and-violet end of the spectrum; the other responds primarily to reddish colors but can also dimly detect blues. The overlapping sensitivity of the two receptors to the blue side of the spectrum allows dogs to distinguish between closely related colors there, because different hues stimulate the two receptors in different proportions. The lack of dual sensitivity to the other side of the spectrum accounts for the dogs' inability to tell hunks of beef and cheese apart (based on color only, that is).

In this way dogs resemble humans who have a partial colorblindness called deuteranopia. Most humans—indeed, most primates—have red-, blue- and green-sensitive photoreceptors, which provide overlapping coverage of the entire spectrum. But about one out of every 100 American males

The bones of a whale can support a community for years on the lightless deep-sea floor



WHALE'S SKELETON (left) rests on the ocean floor in a mosaic of photographs taken from the research submarine *Alvin*. Mussels nestle in an eroded rib bone (above) recovered by the vessel.

is born with a genetic defect that eliminates the green-sensitive receptor. Deuteranopes, like dogs, are sensitive to blues but see greens, yellows, oranges and reds as one shade.

Whence came the myth of total canine colorblindness? According to Jacobs, the most recent investigation into the subject, done in 1969, provided some evidence that dogs see color but was not "compelling." (One problem was that the experiments did not test whether the dogs were responding to differences in brightness rather than color.) Many scientists apparently preferred to believe earlier reports that came to the opposite conclusion. Some textbook writers, extrapolating from these spurious findings, even proclaimed that all nonprimate mammals were colorblind.

Actually, Jacobs says, experiments by his group and others have found evidence—physiological and behavioral—of two-receptor color vision in squirrels, shrews, pigs and cats as well as dogs. Indeed, although a few strongly nocturnal mammals—such as rats and hamsters—have been shown to be completely colorblind, Jacobs speculates that most mammals share to some extent our colorful vision of the world.

—J.H.

MEDICINE

Nervous Excitement

Studies of nerve regeneration take a step forward

Nerves in the brain and spinal cord of mammals do not naturally grow back if they are cut or badly injured. Consequently, people who have suffered injuries to the central nervous system can be permanently disabled. It was observed early in the 1980's, however, that mature damaged neurons could sometimes regrow their axons—the armlike extensions along which they transmit signals—through grafts from peripheral nerves in the limbs, which can often regenerate naturally. The peripheral nerve tissue serves as a bridge for guiding (and perhaps stimulating) the regrowth of axons.

David A. Carter, Garth M. Bray and Albert J. Aguayo of the Montreal General Hospital and the Center for Research in Neuroscience at McGill University have now reported in the *Journal of Neuroscience* that regenerated neurons can form synaptic connections with other neurons that are vir-



COLOR PERCEPTION of dogs was tested at the University of California at Santa Barbara. If Retina, the toy poodle shown above, pushes the blue screen with her nose, she will receive a food pellet in the corresponding dish.

tually identical to those of normal cells. Carter and his colleagues cut the optic nerves of four hamsters, then grafted a three-centimeter length of peripheral nerve between the injured retinal neurons and their normal destination in the brain, an area called the superior colliculus.

After about seven weeks, Carter's group found that the retinal neurons had regrown their axons into the superior colliculus and formed structurally normal synapses with the superior colliculus neurons. Equally important, the regenerated neurons had made their connections within the correct layer of the superior colliculus and to the appropriate structures (the antennalike dendrites rather than the neural cell bodies).

The next question was whether these seemingly normal synapses could pass electrical signals from one neuron to the next; a report appearing in *Science* demonstrates that the answer is yes. Susan A. Keirstead, Michael Rasminsky and Yutaka Fukuda, working in collaboration with Carter, Aguayo and Manuel Vidal-Sanz of McGill, experimented on eight hamsters that had received grafts to reconnect their optic nerves about four months previously. Keirstead and her colleagues flashed a light in each animal's eyes and monitored activity in the superior colliculus with electrodes. They found that excitement of the retina with light did evoke excitatory and inhibitory activity in the superior colliculus neurons, a clear indication that the synapses between the

regenerated axons and their superior colliculus targets were, in at least some respects, functional.

By no means does this suggest that the hamsters had regained their sight. As Rasminsky points out, "We have as yet little information about the number and appropriateness of the reconnections made by the regenerated axons. Obviously, one of the many directions for future research is determining how much specificity there is in the pattern of reconnections."

Mary Ellen Michel, who manages a program studying central nervous system injuries and regeneration at the National Institutes of Health, echoes Rasminsky's caution: "It's a long way from this kind of demonstration to people walking after a spinal cord injury, but it does help keep the hope alive."

—John Rennie

Sleep of the Guilty

The case of the bulimic, herpetophobic somnambulist

Should people be held responsible for crimes they commit while sleeping? The prevailing view among psychiatrists is that they should not, because, as one expert in the field has stated, "the sleeping mind cannot form an intent." In a recent letter to *Lancet*, Peter Roper, a psychiatrist at McGill University, challenges this view. He offers as contrary evidence the case of the bulimic, herpetophobic somnambulist.

A Montreal housewife, she could control her urge to gorge while awake. But at night, Roper writes, "she would sleepwalk to her refrigerator and eat anything that was there—raw meat, butter, uncooked vegetables. If she woke sufficiently to realize what she was doing she would feel disgusted and return to bed—but usually she was unaware of her nocturnal forays until discarded wrappers, food remnants, and general disorder were found in the kitchen in the morning."

On psychoanalyzing the woman, Roper found that she also had an intense fear of snakes. Indeed, she was terrified even of a rubber snake that belonged to one of her two children. Roper decided to pit her phobia against her bulimia. He instructed the woman's husband to set the toy snake on a table near the refrigerator every night before he and his wife went to bed.

The ploy worked. Over the next two and a half years, the woman raided the refrigerator on a total of only six nights. On each of these occasions, the woman confessed to Roper later, she had noticed before retiring that her husband had forgotten to put the snake out; she neglected to remind him of his duty.

On these nights, Roper points out, the sleeping woman exploited knowledge gained while she was awake to satisfy her compulsive desire for food. Clearly, he says, sleepwalkers are not the mindless automatons that some psychiatrists have depicted them as. "The concept that 'the sleeping mind cannot form an intent,'" he concludes, "may therefore be misleading, and doctors should take this into account when giving expert testimony in criminal cases involving sleepwalking."

Such cases are unusual but not unheard of, according to Roper. He notes that several years ago, for example, a man living in the Toronto area claimed he was asleep when he drove 10 miles to the home of his mother-in-law and then beat and stabbed her to death. The jury acquitted him. —J.H.

PROFILE

Claude E. Shannon

Unicyclist, juggler and father of information theory

Claude E. Shannon can't sit still. We're at his home, a stuccoed Victorian edifice overlooking a lake north of Boston, and I'm trying to

get him to recall how he came up with the theory of information. But Shannon, who is a boyish 73, with an elfish grin and a shock of snowy hair, is tired of expounding on his past. Wouldn't I rather see his toys?

Without waiting for an answer, and over the mild protests of his wife, Betty, he leaps from his chair and disappears into the other room. When I catch up with him, he proudly shows me his seven chess-playing machines, gasoline-powered pogostick, hundred-bladed jackknife, two-seated unicycle and countless other marvels. Some of his personal creations—such as a juggling W. C. Fields mannequin and a computer called THROBAC that calculates in Roman numerals—are a bit dusty and in disrepair, but Shannon seems as delighted with everything as a 10-year-old on Christmas morning.

Is this the man who, as a young engineer at Bell Laboratories in 1948, wrote the Magna Carta of the information age: "The Mathematical Theory of Communication"? Whose work Robert W. Lucky, executive director of research at AT&T Bell Laboratories, calls the greatest "in the annals of technological thought"? Whose "pioneering insight" IBM Fellow Rolf W. Landauer equates with Einstein's? Yes. This is also the man who invented a rocket-powered Frisbee and who juggled while riding a unicycle through the halls of Bell Labs. "I've always pursued my interests without much regard to financial value or value to the world," Shannon says. "I've spent lots of time on totally useless things."

From childhood on, Shannon was fascinated by both the particulars of hardware and the generalities of mathematics. Growing up in Gaylord, Mich., he tinkered with erector sets and radios given to him by his father, a probate judge, and solved mathematical puzzles supplied by his older sister, Catherine, who became a professor of mathematics. As an undergraduate at the University of Michigan he majored in electrical engineering and mathematics.

His familiarity with the two fields helped him notch his first big success while he was still a graduate student at the Massachusetts Institute of Technology. In his master's thesis he showed how an algebra invented in the mid-1800's by the British mathematician George Boole—which deals with such concepts as "if X or Y happens and not Z, then Q results"—could represent the workings of switches and relays in electronic circuits.

The implications of the paper were profound: engineers now routinely de-

sign computer hardware and software, telephone networks and other systems with the aid of Boolean algebra. Shannon downplays the discovery. "It just happened that no one else was familiar with both those fields at the same time," he says. He adds, after a moment of reflection, "I've always loved that word, 'Boolean.'"

In 1941, a year after obtaining his Ph.D. from M.I.T., Shannon went to Bell Labs. During World War II his official responsibility was developing cryptographic systems, but on his own time he nurtured the ideas that were to evolve into information theory. Shannon's initial goal was simple: to improve the transmission of information over a telegraph or telephone line affected by electrical interference, or noise. The best solution, he decided, was not to improve transmission lines but to package information more efficiently.

What is information? Sidestepping questions about meaning, Shannon showed that it is a measurable commodity: the amount of information in a given message is a function of the probability that—out of all the messages that could be sent—it would be selected. He defined the overall potential for information in a system of messages as its "entropy," which in thermodynamics denotes the randomness—or "shuffledness," if you will—of a system. (Shannon once said that the great mathematician John von Neumann had urged him to use the term entropy, pointing out that since no one really knows what it means, Shannon would have an advantage in debates about his theory.)

Shannon defined the basic unit of information, which came to be called a bit, as a message representing one of two choices: heads or tails, for example, or yes or no. One could encode great amounts of information in bits, just as in the old game "20 Questions" one could quickly zero in on a correct answer through deft questioning. A bit can be represented as a one or a zero or as the presence or absence of current in a wire.

Building on this mathematical foundation, Shannon then showed that any given communications channel has a maximum capacity for reliably transmitting information. Actually he showed that although one can approach this maximum through clever coding, one can never quite reach it. The maximum has come to be known as the Shannon limit.

How does one approach the Shannon limit? The first step is to eliminate redundancy. Just as a laconic

suitor might write "I lv u" in his billet-doux, so will a good code compress information to its most compact form. The code then adds just enough redundancy to ensure that the stripped-down message is not obscured by noise. For example, a code processing a stream of numbers might add a polynomial equation on whose graph the numbers all fall. The decoder on the receiving end knows that any numbers that diverge from the graph have been altered in transmission.

Shannon's ideas were almost too prescient to have an immediate practical impact. Vacuum-tube circuits simply could not calculate the complex codes needed to approach the Shannon limit. In fact, not until the early 1970's—with the advent of high-speed integrated circuits—did engineers begin fully to exploit information theory. Today Shannon's insights help shape virtually all systems that store, process or transmit information in digital form, from compact disks to computers, from facsimile machines to deep-space probes such as *Voyager*.

Information theory has also infiltrated fields outside of communications, including linguistics, psychology, economics, biology, even the arts. In the early 1970's the *IEEE Transactions on Information Theory* published an editorial, titled "Information Theory, Photosynthesis and Religion," decrying this trend. Yet Shannon himself suggests that applying information theory to biological systems may not be so farfetched, because in his view common principles underlie mechanical and living things. "You bet," he replies, when asked whether he thinks machines can think. "I'm a machine and you're a machine, and we both think, don't we?"

Indeed, Shannon was one of the first engineers to propose that machines could be programmed to play games and perform other complex tasks [see "A Chess-Playing Machine," by Claude E. Shannon; *SCIENTIFIC AMERICAN*, February, 1950]. In 1950 he built Theseus, a mechanical mouse that—guided by a magnet and a mass of circuitry under the floor—could learn how to find its way out of a maze. The invention inspired the Institute of Electrical and Electronics Engineers to initiate a "micromouse" contest in which thousands of engineering students worldwide now participate.

He built a "mind-reading" machine that played the game of penny-matching, in which one person tries to guess whether the other has chosen heads or tails. A colleague at Bell Labs, David W. Hagelbarger, built the prototype;

the machine recorded and analyzed its opponent's past choices, looking for patterns that would foretell the next choice. Because it is almost impossible for a human to avoid falling into such patterns, the machine won more than 50 percent of the time. Shannon then built his own version and challenged Hagelbarger to a legendary duel. Shannon's machine won.

Shannon left Bell Labs to become a professor at M.I.T. in 1956. Since his formal retirement in 1978, his great obsession has been juggling. He has built several juggling machines and devised what may be the unified field theory of juggling: if B equals the number of balls, H the number of hands, D the time each ball spends in a hand, F the time of flight of each ball and E the time each hand is empty, then $B/H = (D + F)/(D + E)$. (Unfortunately, the theory never helped Shannon break his personal record of four balls at once.) Shannon has also developed various mathematical models of the stock market and tested them—successfully, he says—on his own portfolio. He has even dabbled in poetry: among his works is *A Rubric on Rubik Cubics*, set to

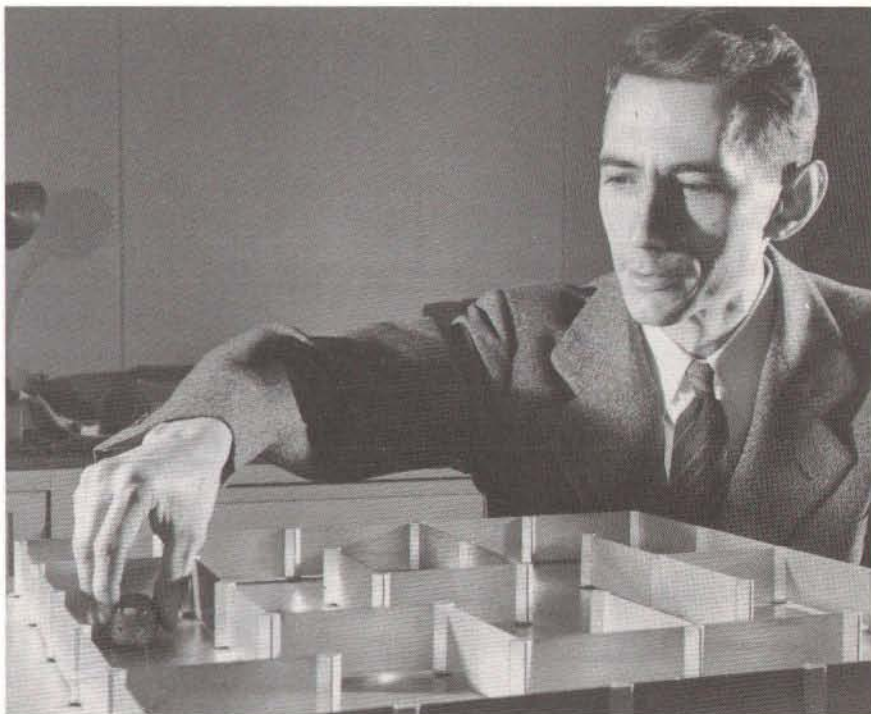
the meter of "Ta-Ra-Ra-Boom-De-Aye."

Shannon has published little on information theory, however, since the late 1950's. Some former Bell colleagues suggest that Shannon had "burned out" and tired of the field he created, but Shannon denies it. He says he continued to study various problems in information theory—at least through the 1960's—but did not consider his investigations good enough to publish. "Most great mathematicians have done their finest work when they were young," he says.

In 1985 Shannon and his wife decided on a whim to visit the International Information Theory Symposium being held in Brighton, England. He had not attended a meeting in many years, and at first no one noticed him. Then word raced around the conference: that shy, white-haired gent wandering in and out of technical sessions was Claude E. Shannon. At the banquet, Shannon said a few words, briefly juggled three balls and then signed autographs for a long line of engineers. Recalls Robert J. McEliece of the California Institute of Technology, "It was as if Newton had showed up at a physics conference."

—John Horgan

Asked if machines think, Shannon answers: You bet. We're machines, and we think, don't we?



THESEUS, a mechanical mouse, is shown with its inventor, Claude E. Shannon, in a 1952 photograph from Bell Laboratories. The mouse "learns" to find its way out of a maze with the help of a magnet guided by circuitry beneath the maze.

Your business can reach the people who make the future happen in France.

POUR LA SCIENCE is the French-language edition of **SCIENTIFIC AMERICAN**. More than 60,000 people buy this prestigious science and technology magazine each month, people who make decisions for France's industry, government and business sectors.

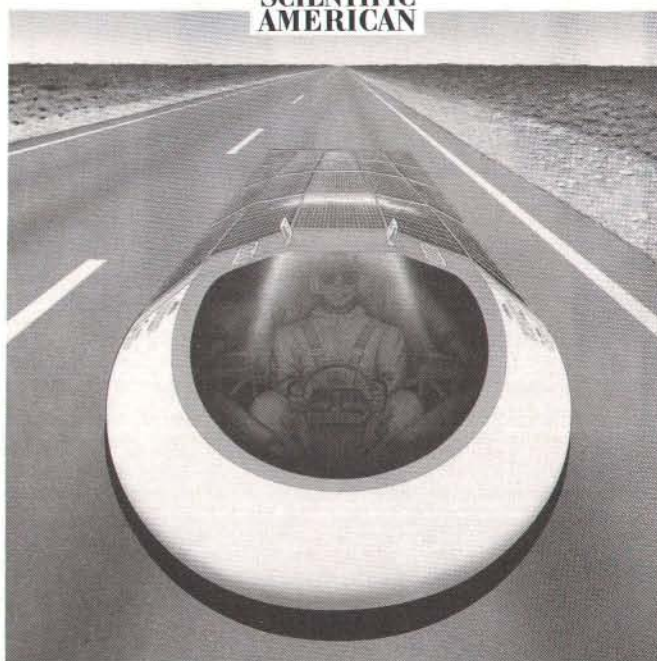
Contact
Susan Mackie at:

POUR LA SCIENCE
8, rue Férou
75006 Paris, France

Telephone
(33) (1) 46-34-21-42
Fax
(33) (1) 43-25-18-29
Telex
842202978

■ POUR LA **SCIENCE**

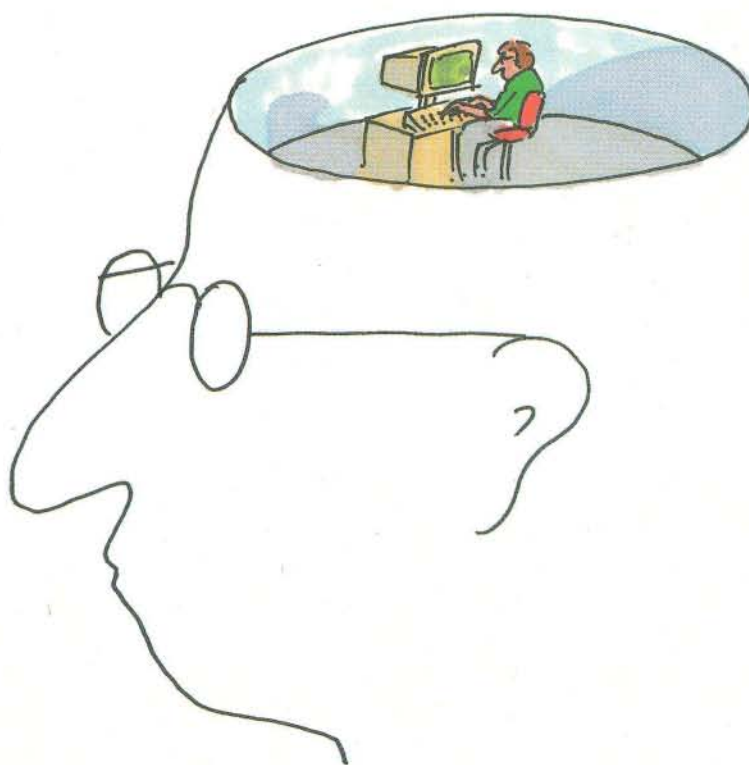
édition française de
**SCIENTIFIC
AMERICAN**



MAI 1989 - MENUEL N° 139
ISSUE 2778 1008 475 CANADA \$3.75 (421) MEXICO 28.00

■ LES VÉHICULES ÉLECTRIQUES ■ LES ACCÉLÉRATEURS À PLASMA
■ ALLIAGES AMORPHES ■ LES VERRES ■ L'AQUEDUC DE NÎMES
■ LA BIOLOGIE DES OBSESSIONS

Artificial Intelligence: A Debate



Atttempts to produce thinking machines have met during the past 35 years with a curious mix of progress and failure. Computers have mastered intellectual tasks such as chess and integral calculus, but they have yet to attain the skills of a lobster in dealing with the real world. Some outside the AI field have argued that the quest is bound to fail: computers by their nature are incapable of true cognition. In the following pages, John R. Searle of the University of California at Berkeley maintains that computer programs can never give rise to minds. On the other side, Paul M. Churchland and Patricia Smith Churchland of the University of California at San Diego claim that circuits modeled on the brain might well achieve intelligence. Behind this debate lies the question, What does it mean to think? The issue has intrigued people (the only entities known to think) for millennia. Computers that so far do not think have given the question a new slant and struck down many candidate answers. A definitive one remains to be found.

Is the Brain's Mind a Computer Program?

No. A program merely manipulates symbols, whereas a brain attaches meaning to them

by John R. Searle

Can a machine think? Can a machine have conscious thoughts in exactly the same sense that you and I have? If by "machine" one means a physical system capable of performing certain functions (and what else can one mean?), then humans are machines of a special biological kind, and humans can think, and so of course machines can think. And, for all we know, it might be possible to produce a thinking machine out of different materials altogether—say, out of silicon chips or vacuum tubes. Maybe it will turn out to be impossible, but we certainly do not know that yet.

In recent decades, however, the question of whether a machine can think has been given a different interpretation entirely. The question that has been posed in its place is, Could a machine think just by virtue of implementing a computer program? Is the program by itself constitutive of thinking? This is a completely different question because it is not about the physical, causal properties of actual or possible physical systems but rather about the abstract, computational properties of formal computer programs that can be implemented in any sort of substance at all, provided only that the substance is able to carry the program.

A fair number of researchers in artificial intelligence (AI) believe the answer to the second question is yes; that is, they believe that by designing the right programs with the right inputs and outputs, they are literally

creating minds. They believe furthermore that they have a scientific test for determining success or failure: the Turing test devised by Alan M. Turing, the founding father of artificial intelligence. The Turing test, as currently understood, is simply this: if a computer can perform in such a way that an expert cannot distinguish its performance from that of a human who has a certain cognitive ability—say, the ability to do addition or to understand Chinese—then the computer also has that ability. So the goal is to design programs that will simulate human cognition in such a way as to pass the Turing test. What is more, such a program would not merely be a model of the mind; it would literally be a mind, in the same sense that a human mind is a mind.

By no means does every worker in artificial intelligence accept so extreme a view. A more cautious approach is to think of computer models as being useful in studying the mind in the same way that they are useful in studying the weather, economics or molecular biology. To distinguish these two approaches, I call the first strong AI and the second weak AI. It is important to see just how bold an approach strong AI is. Strong AI claims that thinking is merely the manipulation of formal symbols, and that is exactly what the computer does: manipulate formal symbols. This view is often summarized by saying, "The mind is to the brain as the program is to the hardware."

Strong AI is unusual among theories of the mind in at least two respects: it can be stated clearly, and it admits of a simple and decisive refutation. The refutation is one that any person can try for himself or herself. Here is how it goes. Consider a language you don't understand. In my case, I do not understand Chinese. To

me Chinese writing looks like so many meaningless squiggles. Now suppose I am placed in a room containing baskets full of Chinese symbols. Suppose also that I am given a rule book in English for matching Chinese symbols with other Chinese symbols. The rules identify the symbols entirely by their shapes and do not require that I understand any of them. The rules might say such things as, "Take a squiggle-squiggle sign from basket number one and put it next to a squoggle-squoggle sign from basket number two."

Imagine that people outside the room who understand Chinese hand in small bunches of symbols and that in response I manipulate the symbols according to the rule book and hand back more small bunches of symbols. Now, the rule book is the "computer program." The people who wrote it are "programmers," and I am the "computer." The baskets full of symbols are the "data base," the small bunches that are handed in to me are "questions" and the bunches I then hand out are "answers."

Now suppose that the rule book is written in such a way that my "answers" to the "questions" are indistinguishable from those of a native Chinese speaker. For example, the people outside might hand me some symbols that unknown to me mean, "What's your favorite color?" and I might after going through the rules give back symbols that, also unknown to me, mean, "My favorite is blue, but I also like green a lot." I satisfy the Turing test for understanding Chinese. All the same, I am totally ignorant of Chinese. And there is no way I could come to understand Chinese in the system as described, since there is no way that I can learn the meanings of any of the symbols. Like a computer, I manipulate symbols, but I attach no meaning to the symbols.

The point of the thought experiment is this: if I do not understand Chinese solely on the basis of running a computer program for understanding Chinese, then neither does any other digital computer solely on that basis. Digital computers merely manipulate formal symbols according to rules in the program.

What goes for Chinese goes for other forms of cognition as well. Just manipulating the symbols is not by itself enough to guarantee cognition, perception, understanding, thinking and so forth. And since computers, qua computers, are symbol-manipulating devices, merely running the computer program is not enough to guarantee cognition.

JOHN R. SEARLE is professor of philosophy at the University of California, Berkeley. He received his B.A., M.A. and D.Phil. from the University of Oxford, where he was a Rhodes scholar. He wishes to thank Stuart Dreyfus, Stevan Harnad, Elizabeth Lloyd and Irvin Rock for their comments and suggestions.

This simple argument is decisive against the claims of strong AI. The first premise of the argument simply states the formal character of a computer program. Programs are defined in terms of symbol manipulations, and the symbols are purely formal, or "syntactic." The formal character of the program, by the way, is what makes computers so powerful. The same program can be run on an indefinite variety of hardware, and one hardware system can run an indefinite range of computer programs. Let me abbreviate this "axiom" as

Axiom 1. Computer programs are formal (syntactic).

This point is so crucial that it is worth explaining in more detail. A digital computer processes information by first encoding it in the symbolism that the computer uses and then manipulating the symbols through a set of precisely stated rules. These rules constitute the program. For example, in Turing's early theory of computers, the symbols were simply 0's and 1's, and the rules of the program said such things as, "Print a 0 on the tape, move one square to the left and erase a 1." The astonishing thing about computers is that any information that can be stated in a language can be encoded in such a system, and any information-processing task that can be solved by explicit rules can be programmed.

Two further points are important. First, symbols and programs are purely abstract notions: they have no essential physical properties to define them and can be implemented in any physical medium whatsoever. The 0's and 1's, qua symbols, have no essential physical properties and a fortiori have no physical, causal properties. I emphasize this point because it is tempting to identify computers with some specific technology—say, silicon chips—and to think that the issues are about the physics of silicon chips or to think that syntax identifies some physical phenomenon that might have as yet unknown causal powers, in the way that actual physical phenomena such as electromagnetic radiation or hydrogen atoms have physical, causal properties. The second point is that symbols are manipulated without reference to any meanings. The symbols of the program can stand for anything the programmer or user wants. In this sense the program has syntax but no semantics.

The next axiom is just a reminder of the obvious fact that thoughts, perceptions, understandings and so forth have a mental content. By virtue of

their content they can be about objects and states of affairs in the world. If the content involves language, there will be syntax in addition to semantics, but linguistic understanding requires at least a semantic framework. If, for example, I am thinking about the last presidential election, certain words will go through my mind, but the words are about the election only because I attach specific meanings to these words, in accordance with my knowledge of English. In this respect they are unlike Chinese symbols for me. Let me abbreviate this axiom as

Axiom 2. Human minds have mental contents (semantics).

Now let me add the point that the Chinese room demonstrated. Having the symbols by themselves—just having the syntax—is not sufficient for having the semantics. Merely manipulating symbols is not enough to guarantee knowledge of what they mean. I shall abbreviate this as

Axiom 3. Syntax by itself is neither constitutive of nor sufficient for semantics.

At one level this principle is true by definition. One might, of course, define the terms syntax and semantics differently. The point is that there is a distinction between formal elements, which have no intrinsic meaning or content, and those phenomena that have intrinsic content. From these premises it follows that

Conclusion 1. Programs are neither constitutive of nor sufficient for minds.

And that is just another way of saying that strong AI is false.

It is important to see what is proved and not proved by this argument.

First, I have not tried to prove that "a computer cannot think." Since anything that can be simulated computationally can be described as a computer, and since our brains can at some levels be simulated, it follows trivially that our brains are computers and they can certainly think. But from the fact that a system can be simulated by symbol manipulation and the fact that it is thinking, it does not follow that thinking is equivalent to formal symbol manipulation.

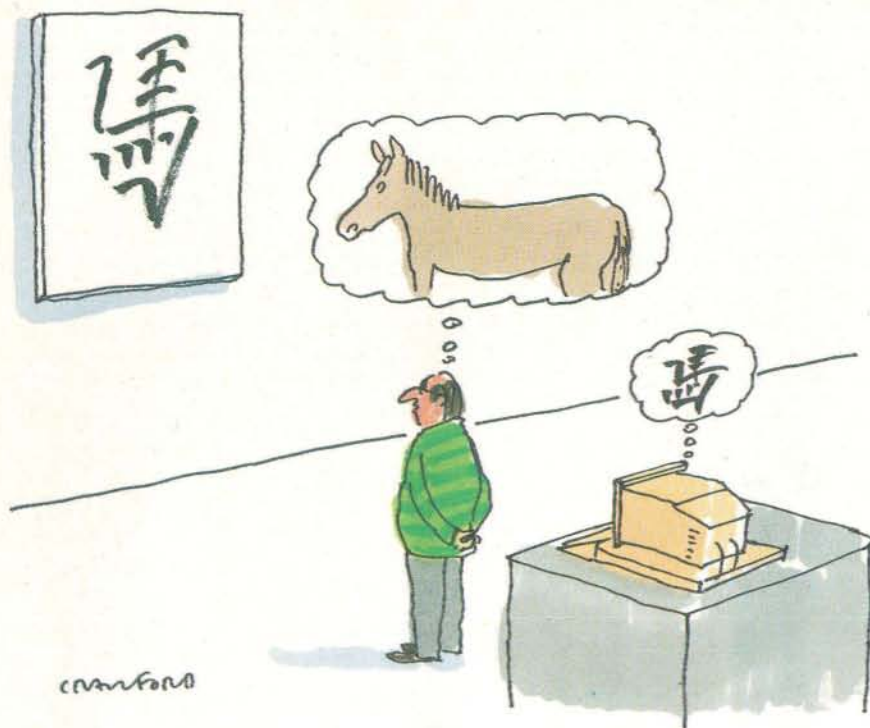
Second, I have not tried to show that only biologically based systems like our brains can think. Right now those are the only systems we know for a fact can think, but we might find other systems in the universe that can produce conscious thoughts, and we might even come to be able to create thinking systems artificially. I regard this issue as up for grabs.

Third, strong AI's thesis is not that, for all we know, computers with the right programs might be thinking, that they might have some as yet undetected psychological properties; rather it is that they must be thinking because that is all there is to thinking.

Fourth, I have tried to refute strong AI so defined. I have tried to demonstrate that the program by itself is not constitutive of thinking because the program is purely a matter of formal symbol manipulation—and we know independently that symbol manipulations by themselves are not sufficient to guarantee the presence of mean-



I satisfy the Turing test for understanding Chinese



*Computer programs are formal (syntactic).
Human minds have mental contents (semantics)*

ings. That is the principle on which the Chinese room argument works.

I emphasize these points here partly because it seems to me the Churchlands [see "Could a Machine Think?" by Paul M. Churchland and Patricia Smith Churchland, page 26] have not quite understood the issues. They think that strong AI is claiming that computers might turn out to think and that I am denying this possibility on commonsense grounds. But that is not the claim of strong AI, and my argument against it has nothing to do with common sense.

I will have more to say about their objections later. Meanwhile I should point out that, contrary to what the Churchlands suggest, the Chinese room argument also refutes any strong-AI claims made for the new parallel technologies that are inspired by and modeled on neural networks. Unlike the traditional von Neumann computer, which proceeds in a step-by-step fashion, these systems have many computational elements that operate in parallel and interact with one another according to rules inspired by neurobiology. Although the results are still modest, these "parallel distributed processing," or "connectionist," models raise useful questions about how complex, parallel network systems like those in brains might actually function in the production of intelligent behavior.

The parallel, "brainlike" character of the processing, however, is irrelevant to the purely computational aspects of the process. Any function that can be computed on a parallel machine can also be computed on a serial machine. Indeed, because parallel machines are still rare, connectionist programs are usually run on traditional serial machines. Parallel processing, then, does not afford a way around the Chinese room argument.

What is more, the connectionist system is subject even on its own terms to a variant of the objection presented by the original Chinese room argument. Imagine that instead of a Chinese room, I have a Chinese gym: a hall containing many monolingual, English-speaking men. These men would carry out the same operations as the nodes and synapses in a connectionist architecture as described by the Churchlands, and the outcome would be the same as having one man manipulate symbols according to a rule book. No one in the gym speaks a word of Chinese, and there is no way for the system as a whole to learn the meanings of any Chinese words. Yet with appropriate adjustments, the system could give the correct answers to Chinese questions.

There are, as I suggested earlier, interesting properties of connectionist nets that enable them to simulate brain processes more accurately than

traditional serial architecture does. But the advantages of parallel architecture for weak AI are quite irrelevant to the issues between the Chinese room argument and strong AI.

The Churchlands miss this point when they say that a big enough Chinese gym might have higher-level mental features that emerge from the size and complexity of the system, just as whole brains have mental features that are not had by individual neurons. That is, of course, a possibility, but it has nothing to do with computation. Computationally, serial and parallel systems are equivalent: any computation that can be done in parallel can be done in serial. If the man in the Chinese room is computationally equivalent to both, then if he does not understand Chinese solely by virtue of doing the computations, neither do they. The Churchlands are correct in saying that the original Chinese room argument was designed with traditional AI in mind but wrong in thinking that connectionism is immune to the argument. It applies to any computational system. You can't get semantically loaded thought contents from formal computations alone, whether they are done in serial or in parallel; that is why the Chinese room argument refutes strong AI in any form.

Many people who are impressed by this argument are nonetheless puzzled about the differences between people and computers. If humans are, at least in a trivial sense, computers, and if humans have a semantics, then why couldn't we give semantics to other computers? Why couldn't we program a Vax or a Cray so that it too would have thoughts and feelings? Or why couldn't some new computer technology overcome the gulf between form and content, between syntax and semantics? What, in fact, are the differences between animal brains and computer systems that enable the Chinese room argument to work against computers but not against brains?

The most obvious difference is that the processes that define something as a computer—computational processes—are completely independent of any reference to a specific type of hardware implementation. One could in principle make a computer out of old beer cans strung together with wires and powered by windmills.

But when it comes to brains, although science is largely ignorant of how brains function to produce mental states, one is struck by the extreme specificity of the anatomy and the

physiology. Where some understanding exists of how brain processes produce mental phenomena—for example, pain, thirst, vision, smell—it is clear that specific neurobiological processes are involved. Thirst, at least of certain kinds, is caused by certain types of neuron firings in the hypothalamus, which in turn are caused by the action of a specific peptide, angiotensin II. The causation is from the “bottom up” in the sense that lower-level neuronal processes cause higher-level mental phenomena. Indeed, as far as we know, every “mental” event, ranging from feelings of thirst to thoughts of mathematical theorems and memories of childhood, is caused by specific neurons firing in specific neural architectures.

But why should this specificity matter? After all, neuron firings could be simulated on computers that had a completely different physics and chemistry from that of the brain. The answer is that the brain does not merely instantiate a formal pattern or program (it does that, too), but it also *causes* mental events by virtue of specific neurobiological processes. Brains are specific biological organs, and their specific biochemical properties enable them to cause consciousness and other sorts of mental phenomena. Computer simulations of brain processes provide models of the formal aspects of these processes. But the simulation should not be confused with duplication. The computational model of mental processes is no more real than the computational model of any other natural phenomenon.

One can imagine a computer simulation of the action of peptides in the hypothalamus that is accurate down to the last synapse. But equally one can imagine a computer simulation of the oxidation of hydrocarbons in a car engine or the action of digestive processes in a stomach when it is digesting pizza. And the simulation is no more the real thing in the case of the brain than it is in the case of the car or the stomach. Barring miracles, you could not run your car by doing a computer simulation of the oxidation of gasoline, and you could not digest pizza by running the program that simulates such digestion. It seems obvious that a simulation of cognition will similarly not produce the effects of the neurobiology of cognition.

All mental phenomena, then, are caused by neurophysiological processes in the brain. Hence,

Axiom 4. Brains cause minds.

In conjunction with my earlier derivation, I immediately derive, trivially,

Conclusion 2. Any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains.

This is like saying that if an electrical engine is to be able to run a car as fast as a gas engine, it must have (at least) an equivalent power output. This conclusion says nothing about the mechanisms. As a matter of fact, cognition is a biological phenomenon: mental states and processes are caused by brain processes. This does not imply that only a biological system could think, but it does imply that any alternative system, whether made of silicon, beer cans or whatever, would have to have the relevant causal capacities equivalent to those of brains. So now I can derive

Conclusion 3. Any artifact that produced mental phenomena, any artificial brain, would have to be able to duplicate the specific causal powers of brains, and it could not do that just by running a formal program.

Furthermore, I can derive an important conclusion about human brains:

Conclusion 4. The way that human brains actually produce mental phenomena cannot be solely by virtue of running a computer program.

I first presented the Chinese room parable in the pages of *Behavioral and Brain Sciences* in 1980, where it appeared, as is the practice of the journal, along with peer commentary, in this case, 26 commentaries. Frankly, I think the point it makes is rather obvious, but to my surprise the publication was followed by a further flood of objections that—more surprisingly—continues to the present day. The Chinese room argument

clearly touched some sensitive nerve.

The thesis of strong AI is that any system whatsoever—whether it is made of beer cans, silicon chips or toilet paper—not only might have thoughts and feelings but *must* have thoughts and feelings, provided only that it implements the right program, with the right inputs and outputs. Now, that is a profoundly antibiological view, and one would think that people in AI would be glad to abandon it. Many of them, especially the younger generation, agree with me, but I am amazed at the number and vehemence of the defenders. Here are some of the common objections.

a. In the Chinese room you really do understand Chinese, even though you don't know it. It is, after all, possible to understand something without knowing that one understands it.

b. You don't understand Chinese, but there is an (unconscious) subsystem in you that does. It is, after all, possible to have unconscious mental states, and there is no reason why your understanding of Chinese should not be wholly unconscious.

c. You don't understand Chinese, but the whole room does. You are like a single neuron in the brain, and just as such a single neuron by itself cannot understand but only contributes to the understanding of the whole system, you don't understand, but the whole system does.

d. Semantics doesn't exist anyway; there is only syntax. It is a kind of prescientific illusion to suppose that there exist in the brain some mysterious “mental contents,” “thought processes” or “semantics.” All that exists in the brain is the same sort of syntactic symbol manipulation that



Which semantics is the system giving off now?

goes on in computers. Nothing more.

e. You are not really running the computer program—you only think you are. Once you have a conscious agent going through the steps of the program, it ceases to be a case of implementing a program at all.

f. Computers would have semantics and not just syntax if their inputs and outputs were put in appropriate causal relation to the rest of the world. Imagine that we put the computer into a robot, attached television cameras to the robot's head, installed transducers connecting the television messages to the computer and had the computer output operate the robot's arms and legs. Then the whole system would have a semantics.

g. If the program simulated the operation of the brain of a Chinese speaker, then it would understand Chinese. Suppose that we simulated the brain of a Chinese person at the level of neurons. Then surely such a system would understand Chinese as well as any Chinese person's brain.

And so on.

All of these arguments share a common feature: they are all inadequate because they fail to come to grips with the actual Chinese room argument. That argument rests on the distinction between the formal symbol manipulation that is done by the computer and the mental contents biologically produced by the brain, a distinction I have abbreviated—I hope not misleadingly—as the distinction between syntax and semantics. I will not repeat my answers to all of these objections, but it will help to clarify the issues if I explain the weaknesses of the most widely held objection, argument c—what I call the systems reply. (The brain simulator reply, argument g, is another popular one, but I have already addressed that one in the previous section.)

The systems reply asserts that of course *you* don't understand Chinese but the whole system—you, the room, the rule book, the bushel baskets full of symbols—does. When I first heard this explanation, I asked one of its proponents, "Do you mean the room understands Chinese?" His answer was yes. It is a daring move, but aside from its implausibility, it will not work on purely logical grounds. The point of the original argument was that symbol shuffling by itself does not give any access to the meanings of the symbols. But this is as much true of the whole room as it is of the person inside. One can see this point by extending

the thought experiment. Imagine that I memorize the contents of the baskets and the rule book, and I do all the calculations in my head. You can even imagine that I work out in the open. There is nothing in the "system" that is not in me, and since I don't understand Chinese, neither does the system.

The Churchlands in their companion piece produce a variant of the systems reply by imagining an amusing analogy. Suppose that someone said that light could not be electromagnetic because if you shake a bar magnet in a dark room, the system still will not give off visible light. Now, the Churchlands ask, is not the Chinese room argument just like that? Does it not merely say that if you shake Chinese symbols in a semantically dark room, they will not give off the light of Chinese understanding? But just as later investigation showed

that light was entirely constituted by electromagnetic radiation, could not later investigation also show that semantics are entirely constituted of syntax? Is this not a question for further scientific investigation?

Arguments from analogy are notoriously weak, because before one can make the argument work, one has to establish that the two cases are truly analogous. And here I think they are not. The account of light in terms of electromagnetic radiation is a causal story right down to the ground. It is a causal account of the physics of electromagnetic radiation. But the analogy with formal symbols fails because formal symbols have no physical, causal powers. The only power that symbols have, *qua* symbols, is the power to cause the next step in the program when the machine is running. And there is no question of waiting on further research to reveal the physical,



How could anyone have supposed that a computer simulation of a mental process must be the real thing?

causal properties of 0's and 1's. The only relevant properties of 0's and 1's are abstract computational properties, and they are already well known.

The Churchlands complain that I am "begging the question" when I say that uninterpreted formal symbols are not identical to mental contents. Well, I certainly did not spend much time arguing for it, because I take it as a logical truth. As with any logical truth, one can quickly see that it is true, because one gets inconsistencies if one tries to imagine the converse. So let us try it. Suppose that in the Chinese room some undetectable Chinese thinking really is going on. What exactly is supposed to make the manipulation of the syntactic elements into specifically Chinese thought contents? Well, after all, I am assuming that the programmers were Chinese speakers, programming the system to process Chinese information.

Fine. But now imagine that as I am sitting in the Chinese room shuffling the Chinese symbols, I get bored with just shuffling the—to me—meaningless symbols. So, suppose that I decide to interpret the symbols as standing for moves in a chess game. Which semantics is the system giving off now? Is it giving off a Chinese semantics or a chess semantics, or both simultaneously? Suppose there is a third person looking in through the window, and she decides that the symbol manipulations can all be interpreted as stock-market predictions. And so on. There is no limit to the number of semantic interpretations that can be assigned to the symbols because, to repeat, the symbols are purely formal. They have no intrinsic semantics.

Is there any way to rescue the Churchlands' analogy from incoherence? I said above that formal symbols do not have causal properties. But of course the program will always be implemented in some hardware or another, and the hardware will have specific physical, causal powers. And any real computer will give off various phenomena. My computers, for example, give off heat, and they make a humming noise and sometimes crunching sounds. So is there some logically compelling reason why they could not also give off consciousness? No. Scientifically, the idea is out of the question, but it is not something the Chinese room argument is supposed to refute, and it is not something that an adherent of strong AI would wish to defend, because any such giving off would have to derive from the physical features of the implementing medium. But the basic premise of strong

AI is that the physical features of the implementing medium are totally irrelevant. What matters are programs, and programs are purely formal.

The Churchlands' analogy between syntax and electromagnetism, then, is confronted with a dilemma; either the syntax is construed purely formally in terms of its abstract mathematical properties, or it is not. If it is, then the analogy breaks down, because syntax so construed has no physical powers and hence no physical, causal powers. If, on the other hand, one is supposed to think in terms of the physics of the implementing medium, then there is indeed an analogy, but it is not one that is relevant to strong AI.

Because the points I have been making are rather obvious—syntax is not the same as semantics, brain processes cause mental phenomena—the question arises, How did we get into this mess? How could anyone have supposed that a computer simulation of a mental process must be the real thing? After all, the whole point of models is that they contain only certain features of the modeled domain and leave out the rest. No one expects to get wet in a pool filled with Ping-Pong-ball models of water molecules. So why would anyone think a computer model of thought processes would actually think?

Part of the answer is that people have inherited a residue of behaviorist psychological theories of the past generation. The Turing test enshrines the temptation to think that if something behaves as if it had certain mental processes, then it must actually have those mental processes. And this is part of the behaviorists' mistaken assumption that in order to be scientific, psychology must confine its study to externally observable behavior. Paradoxically, this residual behaviorism is tied to a residual dualism. Nobody thinks that a computer simulation of digestion would actually digest anything, but where cognition is concerned, people are willing to believe in such a miracle because they fail to recognize that the mind is just as much a biological phenomenon as digestion. The mind, they suppose, is something formal and abstract, not a part of the wet and slimy stuff in our heads. The polemical literature in AI usually contains attacks on something the authors call dualism, but what they fail to see is that they themselves display dualism in a strong form, for unless one accepts the idea that the mind is completely independent of the brain or of any other physically

specific system, one could not possibly hope to create minds just by designing programs.

Historically, scientific developments in the West that have treated humans as just a part of the ordinary physical, biological order have often been opposed by various rearward actions. Copernicus and Galileo were opposed because they denied that the earth was the center of the universe; Darwin was opposed because he claimed that humans had descended from the lower animals. It is best to see strong AI as one of the last gasps of this antiscientific tradition, for it denies that there is anything essentially physical and biological about the human mind. The mind according to strong AI is independent of the brain. It is a computer program and as such has no essential connection to any specific hardware.

Many people who have doubts about the psychological significance of AI think that computers might be able to understand Chinese and think about numbers but cannot do the crucially human things, namely—and then follows their favorite human specialty—falling in love, having a sense of humor, feeling the angst of postindustrial society under late capitalism, or whatever. But workers in AI complain—correctly—that this is a case of moving the goalposts. As soon as an AI simulation succeeds, it ceases to be of psychological importance. In this debate both sides fail to see the distinction between simulation and duplication. As far as simulation is concerned, there is no difficulty in programming my computer so that it prints out, "I love you, Suzy"; "Ha ha"; or "I am suffering the angst of postindustrial society under late capitalism." The important point is that simulation is not the same as duplication, and that fact holds as much import for thinking about arithmetic as it does for feeling angst. The point is not that the computer gets only to the 40-yard line and not all the way to the goal line. The computer doesn't even get started. It is not playing that game.

FURTHER READING

MIND DESIGN: PHILOSOPHY, PSYCHOLOGY, ARTIFICIAL INTELLIGENCE. Edited by John Haugeland. The MIT Press, 1980.
MINDS, BRAINS, AND PROGRAMS. John Searle in *Behavioral and Brain Sciences*, Vol. 3, No. 3, pages 417-458; 1980.
MINDS, BRAINS, AND SCIENCE. John R. Searle. Harvard University Press, 1984.
MINDS, MACHINES AND SEARLE. Stevan Harnad in *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 1, No. 1, pages 5-25; 1989.

Could a Machine Think?

Classical AI is unlikely to yield conscious machines; systems that mimic the brain might

by Paul M. Churchland and Patricia Smith Churchland

Artificial-intelligence research is undergoing a revolution. To explain how and why, and to put John R. Searle's argument in perspective, we first need a flashback.

By the early 1950's the old, vague question, Could a machine think? had been replaced by the more approachable question, Could a machine that manipulated physical symbols according to structure-sensitive rules think? This question was an improvement because formal logic and computational theory had seen major developments in the preceding half-century. Theorists had come to appreciate the enormous power of abstract systems of symbols that undergo rule-governed transformations. If those systems could just be automated, then their abstract computational power, it seemed, would be displayed in a real physical system. This insight spawned a well-defined research program with deep theoretical underpinnings.

Could a machine think? There were many reasons for saying yes. One of the earliest and deepest reasons lay in

two important results in computational theory. The first was Church's thesis, which states that every effectively computable function is recursively computable. Effectively computable means that there is a "rote" procedure for determining, in finite time, the output of the function for a given input. Recursively computable means more specifically that there is a finite set of operations that can be applied to a given input, and then applied again and again to the successive results of such applications, to yield the function's output in finite time. The notion of a rote procedure is nonformal and intuitive; thus, Church's thesis does not admit of a formal proof. But it does go to the heart of what it is to compute, and many lines of evidence converge in supporting it.

The second important result was Alan M. Turing's demonstration that any recursively computable function can be computed in finite time by a maximally simple sort of symbol-manipulating machine that has come to be called a universal Turing machine. This machine is guided by a set of recursively applicable rules that are sensitive to the identity, order and arrangement of the elementary symbols it encounters as input.

These two results entail something remarkable, namely that a standard digital computer, given only the right program, a large enough memory and sufficient time, can compute *any* rule-governed input-output function. That is, it can display any systematic pattern of responses to the environment whatsoever.

More specifically, these results imply that a suitably programmed symbol-manipulating machine (hereafter, SM machine) should be able to pass the Turing test for conscious intelligence. The Turing test is a purely behavioral test for conscious intelligence, but it is a very demanding test even so. (Whether it is a fair test will be addressed below, where we shall also encounter a second and quite different "test" for conscious in-

telligence.) In the original version of the Turing test, the inputs to the SM machine are conversational questions and remarks typed into a console by you or me, and the outputs are typewritten responses from the SM machine. The machine passes this test for conscious intelligence if its responses cannot be discriminated from the typewritten responses of a real, intelligent person. Of course, at present no one knows the function that would produce the output behavior of a conscious person. But the Church and Turing results assure us that, whatever that (presumably effective) function might be, a suitable SM machine could compute it.

This is a significant conclusion, especially since Turing's portrayal of a purely teletyped interaction is an unnecessary restriction. The same conclusion follows even if the SM machine interacts with the world in more complex ways: by direct vision, real speech and so forth. After all, a more complex recursive function is still Turing-computable. The only remaining problem is to identify the undoubtedly complex function that governs the human pattern of response to the environment and then write the program (the set of recursively applicable rules) by which the SM machine will compute it. These goals form the fundamental research program of classical AI.

Initial results were positive. SM machines with clever programs performed a variety of ostensibly cognitive activities. They responded to complex instructions, solved complex arithmetic, algebraic and tactical problems, played checkers and chess, proved theorems and engaged in simple dialogue. Performance continued to improve with the appearance of larger memories and faster machines and with the use of longer and more cunning programs. Classical, or "program-writing," AI was a vigorous and successful research effort from almost every perspective. The occasional denial that an SM machine might eventually think appeared uninformed and ill motivated. The case for a positive answer to our title question was overwhelming.

There were a few puzzles, of course. For one thing, SM machines were admittedly not very brainlike. Even here, however, the classical approach had a convincing answer. First, the physical material of any SM machine has nothing essential to do with what function it computes. That is fixed by its program. Second, the engineering details of any machine's functional architecture are also irrelevant, since different

PAUL M. CHURCHLAND and PATRICIA SMITH CHURCHLAND are professors of philosophy at the University of California at San Diego. Together they have studied the nature of the mind and knowledge for the past two decades. Paul Churchland focuses on the nature of scientific knowledge and its development, while Patricia Churchland focuses on the neurosciences and on how the brain sustains cognition. Paul Churchland's *Matter and Consciousness* is the standard textbook on the philosophy of the mind, and Patricia Churchland's *Neurophilosophy* brings together theories of cognition from both philosophy and biology. Paul Churchland is currently chair of the philosophy department at UCSD, and the two are, respectively, president and past president of the Society for Philosophy and Psychology. Patricia Churchland is also an adjunct professor at the Salk Institute for Biological Studies in San Diego. The Churchlands are also members of the UCSD cognitive science faculty, its Institute for Neural Computation and its Science Studies program.

architectures running quite different programs can still be computing the same input-output function.

Accordingly, AI sought to find the input-output *function* characteristic of intelligence and the most efficient of the many possible programs for computing it. The idiosyncratic way in which the brain computes the function just doesn't matter, it was said. This completes the rationale for classical AI and for a positive answer to our title question.

Could a machine think? There were also some arguments for saying no. Through the 1960's interesting negative arguments were relatively rare. The objection was occasionally made that thinking was a nonphysical process in an immaterial soul. But such dualistic resistance was neither evolutionarily nor explanatorily plausible. It had a negligible impact on AI research.

A quite different line of objection was more successful in gaining the AI community's attention. In 1972 Hubert L. Dreyfus published a book that was highly critical of the parade-case simulations of cognitive activity. He argued for their inadequacy as simulations of genuine cognition, and he pointed to a pattern of failure in these attempts. What they were missing, he suggested, was the vast store of inarticulate background knowledge every person possesses and the common-sense capacity for drawing on relevant aspects of that knowledge as changing circumstance demands. Dreyfus did not deny the possibility that an artificial physical system of some kind might think, but he was highly critical of the idea that this could be achieved solely by symbol manipulation at the hands of recursively applicable rules.

Dreyfus's complaints were broadly perceived within the AI community, and within the discipline of philosophy as well, as shortsighted and unsympathetic, as harping on the inevitable simplifications of a research effort still in its youth. These deficits might be real, but surely they were temporary. Bigger machines and better programs should repair them in due course. Time, it was felt, was on AI's side. Here again the impact on research was negligible.

Time was on Dreyfus's side as well: the rate of cognitive return on increasing speed and memory began to slacken in the late 1970's and early 1980's. The simulation of object recognition in the visual system, for example, proved computationally intensive to an unexpected degree. Realistic

results required longer and longer periods of computer time, periods far in excess of what a real visual system requires. This relative slowness of the simulations was darkly curious; signal propagation in a computer is roughly a million times faster than in the brain, and the clock frequency of a computer's central processor is greater than any frequency found in the brain by a similarly dramatic margin. And yet, on realistic problems, the tortoise easily outran the hare.

Furthermore, realistic performance

required that the computer program have access to an extremely large knowledge base. Constructing the relevant knowledge base was problem enough, and it was compounded by the problem of how to access just the contextually relevant parts of that knowledge base in real time. As the knowledge base got bigger and better, the access problem got worse. Exhaustive search took too much time, and heuristics for relevance did poorly. Worries of the sort Dreyfus had raised finally began to take hold here

THE CHINESE ROOM

Axiom 1. Computer programs are formal (syntactic).

Axiom 2. Human minds have mental contents (semantics).

Axiom 3. Syntax by itself is neither constitutive of nor sufficient for semantics.

Conclusion 1. Programs are neither constitutive of nor sufficient for minds.

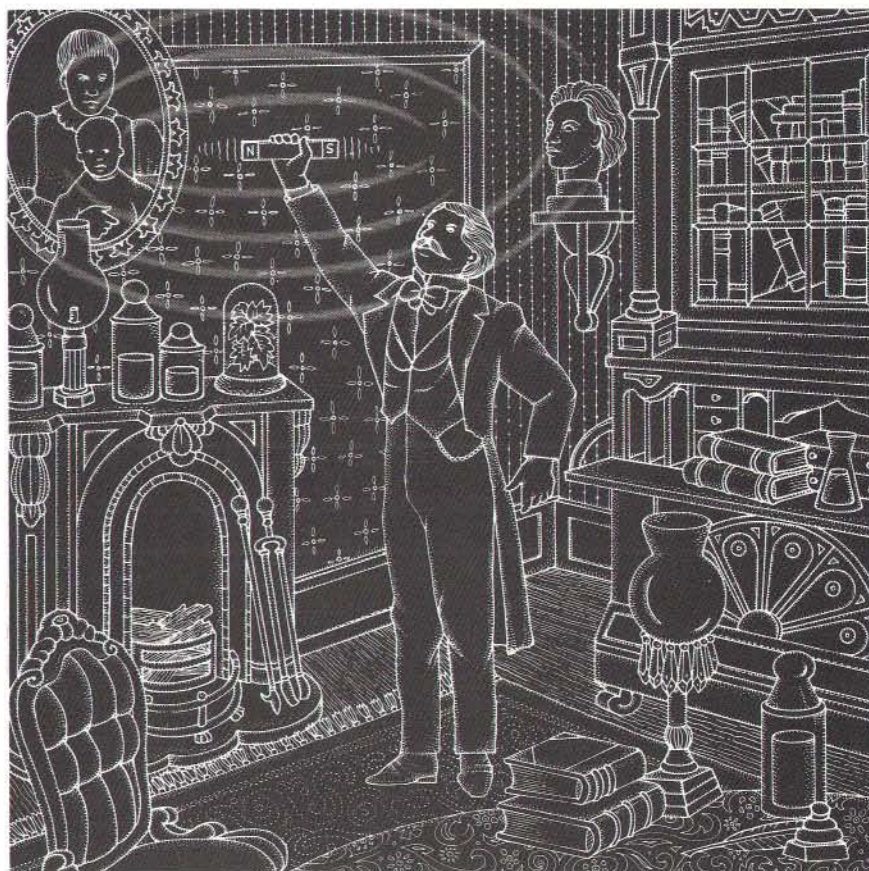
THE LUMINOUS ROOM

Axiom 1. Electricity and magnetism are forces.

Axiom 2. The essential property of light is luminance.

Axiom 3. Forces by themselves are neither constitutive of nor sufficient for luminance.

Conclusion 1. Electricity and magnetism are neither constitutive of nor sufficient for light.



OSCILLATING ELECTROMAGNETIC FORCES constitute light even though a magnet pumped by a person appears to produce no light whatsoever. Similarly, rule-based symbol manipulation might constitute intelligence even though the rule-based system inside John R. Searle's "Chinese room" appears to lack real understanding.

and there even among AI researchers.

At about this time (1980) John Searle authored a new and quite different criticism aimed at the most basic assumption of the classical research program: the idea that the appropriate manipulation of structured symbols by the recursive application of structure-sensitive rules could constitute conscious intelligence.

Searle's argument is based on a thought experiment that displays two crucial features. First, he describes a SM machine that realizes, we are to suppose, an input-output function adequate to sustain a successful Turing test conversation conducted entirely in Chinese. Second, the internal structure of the machine is such that, however it behaves, an observer remains certain that neither the machine nor any part of it understands Chinese. All it contains is a monolingual English speaker following a written set of instructions for manipulating the Chinese symbols that arrive and leave through a mail slot. In short, the system is supposed to pass the Turing test, while the system itself lacks any genuine understanding of Chinese or real Chinese semantic content [see "Is the Brain's Mind a Computer Program?" by John R. Searle, page 20].

The general lesson drawn is that any system that merely manipulates physical symbols in accordance with structure-sensitive rules will be at best a hollow mock-up of real conscious intelligence, because it is impossible to generate "real semantics" merely by cranking away on "empty syntax." Here, we should point out, Searle is imposing a nonbehavioral test for consciousness: the elements of conscious intelligence must possess real semantic content.

One is tempted to complain that Searle's thought experiment is unfair because his Rube Goldberg system will compute with absurd slowness. Searle insists, however, that speed is strictly irrelevant here. A slow thinker should still be a real thinker. Everything essential to the duplication of thought, as per classical AI, is said to be present in the Chinese room.

Searle's paper provoked a lively reaction from AI researchers, psychologists and philosophers alike. On the whole, however, he was met with an even more hostile reception than Dreyfus had experienced. In his companion piece in this issue, Searle forthrightly lists a number of these critical responses. We think many of them are reasonable, especially those that "bite the bullet" by insisting that, although it is appallingly slow, the overall sys-

tem of the room-plus-contents does understand Chinese.

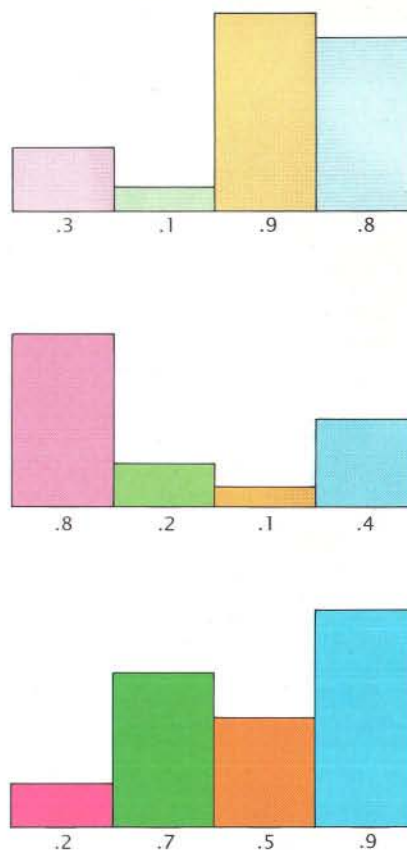
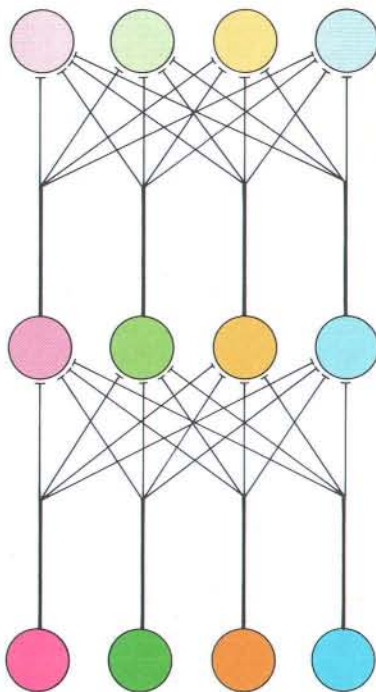
We think those are good responses, but not because we think that the room understands Chinese. We agree with Searle that it does not. Rather they are good responses because they reflect a refusal to accept the crucial third axiom of Searle's argument: "*Syntax by itself is neither constitutive of nor sufficient for semantics.*" Perhaps this axiom is true, but Searle cannot rightly pretend to know that it is. Moreover, to assume its truth is tantamount to begging the question against the research program of classical AI, for that program is predicated on the very interesting assumption that if one can just set in motion an appropriately structured internal dance of syntactic elements, appropriately connected to inputs and outputs, it can produce the same cognitive states and achievements found in human beings.

The question-begging character of Searle's axiom 3 becomes clear when it is compared directly with his con-

clusion 1: "*Programs are neither constitutive of nor sufficient for minds.*" Plainly, his third axiom is already carrying 90 percent of the weight of this almost identical conclusion. That is why Searle's thought experiment is devoted to shoring up axiom 3 specifically. That is the point of the Chinese room.

Although the story of the Chinese room makes axiom 3 tempting to the unwary, we do not think it succeeds in establishing axiom 3, and we offer a parallel argument below in illustration of its failure. A single transparently fallacious instance of a disputed argument often provides far more insight than a book full of logic chopping.

Searle's style of skepticism has ample precedent in the history of science. The 18th-century Irish bishop George Berkeley found it unintelligible that compression waves in the air, by themselves, could constitute or be sufficient for objective sound. The English poet-artist William Blake and the German poet-naturalist Johann W.



NEURAL NETWORKS model a central feature of the brain's microstructure. In this three-layer net, input neurons (bottom left) process a pattern of activations (bottom right) and pass it along weighted connections to a hidden layer. Elements in the hidden layer sum their many inputs to produce a new pattern of activations. This is passed to the output layer, which performs a further transformation. Overall the network transforms any input pattern into a corresponding output pattern as dictated by the arrangement and strength of the many connections between neurons.

von Goethe found it inconceivable that small particles by themselves could constitute or be sufficient for the objective phenomenon of light. Even in this century, there have been people who found it beyond imagining that inanimate matter by itself, and however organized, could ever constitute or be sufficient for life. Plainly, what people can or cannot imagine often has nothing to do with what is or is not the case, even where the people involved are highly intelligent.

To see how this lesson applies to Searle's case, consider a deliberately manufactured parallel to his argument and its supporting thought experiment.

Axiom 1. *Electricity and magnetism are forces.*

Axiom 2. *The essential property of light is luminance.*

Axiom 3. *Forces by themselves are neither constitutive of nor sufficient for luminance.*

Conclusion 1. *Electricity and magnetism are neither constitutive of nor sufficient for light.*

Imagine this argument raised shortly after James Clerk Maxwell's 1864 suggestion that light and electromagnetic waves are identical but before the world's full appreciation of the systematic parallels between the properties of light and the properties of electromagnetic waves. This argument could have served as a compelling objection to Maxwell's imaginative hypothesis, especially if it were accompanied by the following commentary in support of axiom 3.

"Consider a dark room containing a man holding a bar magnet or charged object. If the man pumps the magnet up and down, then, according to Maxwell's theory of artificial luminance (AL), it will initiate a spreading circle of electromagnetic waves and will thus be luminous. But as all of us who have toyed with magnets or charged balls well know, their forces (or any other forces for that matter), even when set in motion, produce no luminance at all. It is inconceivable that you might constitute real luminance just by moving forces around!"

How should Maxwell respond to this challenge? He might begin by insisting that the "luminous room" experiment is a misleading display of the phenomenon of luminance because the frequency of oscillation of the magnet is absurdly low, too low by a factor of 10^{15} . This might well elicit the impatient response that frequency has nothing to do with it, that the room with the bobbing magnet already contains everything essential to light,

according to Maxwell's own theory.

In response Maxwell might bite the bullet and claim, quite correctly, that the room really is bathed in luminance, albeit a grade or quality too feeble to appreciate. (Given the low frequency with which the man can oscillate the magnet, the wavelength of the electromagnetic waves produced is far too long and their intensity is much too weak for human retinas to respond to them.) But in the climate of understanding here contemplated—the 1860's—this tactic is likely to elicit laughter and hoots of derision. "Luminous room, my foot, Mr. Maxwell. It's pitch-black in there!"

Alas, poor Maxwell has no easy route out of this predicament. All he can do is insist on the following three points. First, axiom 3 of the above argument is false. Indeed, it begs the question despite its intuitive plausibility. Second, the luminous room experiment demonstrates nothing of interest one way or the other about the nature of light. And third, what is needed to settle the problem of light and the possibility of artificial luminance is an ongoing research program to determine whether under the appropriate conditions the behavior of electromagnetic waves does indeed mirror perfectly the behavior of light.

This is also the response that classical AI should give to Searle's argument. Even though Searle's Chinese room may appear to be "semantically dark," he is in no position to insist, on the strength of this appearance, that rule-governed symbol manipulation can never constitute semantic phenomena, especially when people have only an uninformed common-sense understanding of the semantic and cognitive phenomena that need to be explained. Rather than exploit one's understanding of these things, Searle's argument freely exploits one's ignorance of them.

With these criticisms of Searle's argument in place, we return to the question of whether the research program of classical AI has a realistic chance of solving the problem of conscious intelligence and of producing a machine that thinks. We believe that the prospects are poor, but we rest this opinion on reasons very different from Searle's. Our reasons derive from the specific performance failures of the classical research program in AI and from a variety of lessons learned from the biological brain and a new class of computational models inspired by its structure. We have already indicated some of the failures of classical AI regarding tasks that the

brain performs swiftly and efficiently. The emerging consensus on these failures is that the functional architecture of classical SM machines is simply the wrong architecture for the very demanding jobs required.

What we need to know is this: How does the brain achieve cognition? Reverse engineering is a common practice in industry. When a new piece of technology comes on the market, competitors find out how it works by taking it apart and divining its structural rationale. In the case of the brain, this strategy presents an unusually stiff challenge, for the brain is the most complicated and sophisticated thing on the planet. Even so, the neurosciences have revealed much about the brain on a wide variety of structural levels. Three anatomic points will provide a basic contrast with the architecture of conventional electronic computers.

First, nervous systems are parallel machines, in the sense that signals are processed in millions of different pathways simultaneously. The retina, for example, presents its complex input to the brain not in chunks of eight, 16 or 32 elements, as in a desktop computer, but rather in the form of almost a million distinct signal elements arriving simultaneously at the target of the optic nerve (the lateral geniculate nucleus), there to be processed collectively, simultaneously and in one fell swoop. Second, the brain's basic processing unit, the neuron, is comparatively simple. Furthermore, its response to incoming signals is analog, not digital, inasmuch as its output spiking frequency varies continuously with its input signals. Third, in the brain, axons projecting from one neuronal population to another are often matched by axons returning from their target population. These descending or recurrent projections allow the brain to modulate the character of its sensory processing. More important still, their existence makes the brain a genuine dynamical system whose continuing behavior is both highly complex and to some degree independent of its peripheral stimuli.

Highly simplified model networks have been useful in suggesting how real neural networks might work and in revealing the computational properties of parallel architectures. For example, consider a three-layer model consisting of neuronlike units fully connected by axonlike connections to the units at the next layer. An input stimulus produces some activation level in a given input unit, which con-

veys a signal of proportional strength along its "axon" to its many "synaptic" connections to the hidden units. The global effect is that a pattern of activations across the set of input units produces a distinct pattern of activations across the set of hidden units.

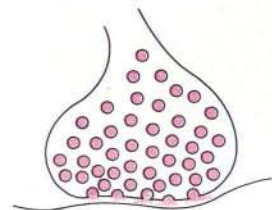
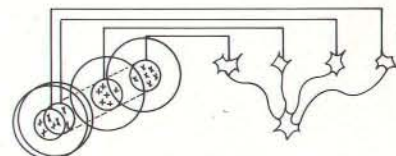
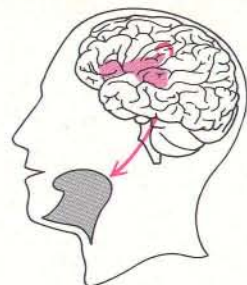
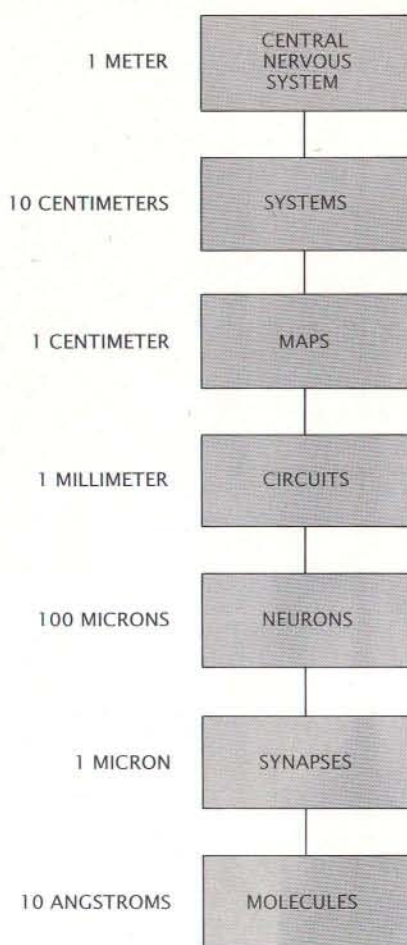
The same story applies to the output units. As before, an activation pattern across the hidden units produces a distinct activation pattern across the output units. All told, this network is a device for transforming any one of a great many possible input vectors (activation patterns) into a uniquely corresponding output vector. It is a device for computing a specific function. Exactly which function it computes is fixed by the global configuration of its synaptic weights.

There are various procedures for adjusting the weights so as to yield a network that computes almost any function—that is, any vector-to-vector transformation—that one might desire. In fact, one can even impose on it a function one is unable to specify, so long as one can supply a set of examples of the desired input-output pairs. This process, called "training up the network," proceeds by successive adjustment of the network's weights until it performs the input-output transformations desired.

Although this model network vastly oversimplifies the structure of the brain, it does illustrate several important ideas. First, a parallel architecture provides a dramatic speed advantage over a conventional computer, for the many synapses at each level perform many small computations simultaneously instead of in laborious sequence. This advantage gets larger as the number of neurons increases at each layer. Strikingly, the speed of processing is entirely independent of both the number of units involved in each layer and the complexity of the function they are computing. Each layer could have four units or a hundred million; its configuration of synaptic weights could be computing simple one-digit sums or second-order differential equations. It would make no difference. The computation time would be exactly the same.

Second, massive parallelism means that the system is fault-tolerant and functionally persistent; the loss of a few connections, even quite a few, has a negligible effect on the character of the overall transformation performed by the surviving network.

Third, a parallel system stores large amounts of information in a distributed fashion, any part of which can be accessed in milliseconds. That in-



NERVOUS SYSTEMS span many scales of organization, from neurotransmitter molecules (*bottom*) to the entire brain and spinal cord. Intermediate levels include single neurons and circuits made up of a few neurons, such as those that produce orientation selectivity to a visual stimulus (*middle*), and systems made up of circuits such as those that subserve language (*top right*). Only research can decide how closely an artificial system must mimic the biological one to be capable of intelligence.

formation is stored in the specific configuration of synaptic connection strengths, as shaped by past learning. Relevant information is "released" as the input vector passes through—and is transformed by—that configuration of connections.

Parallel processing is not ideal for all types of computation. On tasks that require only a small input vector, but many millions of swiftly iterated recursive computations, the brain performs very badly, whereas classical SM machines excel. This class of computations is very large and important, so classical machines will always be useful, indeed, vital. There is, however, an equally large class of computations for which the brain's architecture is the superior technology. These are the computations that typically confront living creatures: recognizing a predator's outline in a noisy environment; recalling instantly how to avoid its gaze, flee its approach or fend

off its attack; distinguishing food from nonfood and mates from non-mates; navigating through a complex and ever-changing physical/social environment; and so on.

Finally, it is important to note that the parallel system described is not manipulating symbols according to structure-sensitive rules. Rather symbol manipulation appears to be just one of many cognitive skills that a network may or may not learn to display. Rule-governed symbol manipulation is not its basic mode of operation. Searle's argument is directed against rule-governed SM machines; vector transformers of the kind we describe are therefore not threatened by his Chinese room argument even if it were sound, which we have found independent reason to doubt.

Searle is aware of parallel processors but thinks they too will be devoid of real semantic content. To illustrate their inevitable failure, he outlines a

second thought experiment, the Chinese gym, which has a gymnasium full of people organized into a parallel network. From there his argument proceeds as in the Chinese room.

We find this second story far less responsive or compelling than his first. For one, it is irrelevant that no unit in his system understands Chinese, since the same is true of nervous systems: no neuron in my brain understands English, although my whole brain does. For another, Searle neglects to mention that his simulation (using one person per neuron, plus a fleet-footed child for each synaptic connection) will require at least 10^{14} people, since the human brain has 10^{11} neurons, each of which averages over 10^3 connections. His system will require the entire human populations of over 10,000 earths. One gymnasium will not begin to hold a fair simulation.

On the other hand, if such a system were to be assembled on a suitably cosmic scale, with all its pathways faithfully modeled on the human case, we might then have a large, slow, oddly made but still functional brain on our hands. In that case the default assumption is surely that, given proper inputs, it would think, not that it couldn't. There is no guarantee that its activity would constitute real thought, because the vector-processing theory sketched above may not be the correct theory of how brains work. But neither is there any a priori guarantee that it could not be thinking. Searle is once more mistaking the limits on his (or the reader's) current imagination for the limits on objective reality.

The brain is a kind of computer, although most of its properties remain to be discovered. Characterizing the brain as a kind of computer is neither trivial nor frivolous. The brain does compute functions, functions of great complexity, but not in the classical AI fashion. When brains are said to be computers, it should not be implied that they are serial, digital computers, that they are programmed, that they exhibit the distinction between hardware and software or that they must be symbol manipulators or rule followers. Brains are computers in a radically different style.

How the brain manages meaning is still unknown, but it is clear that the problem reaches beyond language use and beyond humans. A small mound of fresh dirt signifies to a person, and also to coyotes, that a gopher is around; an echo with a certain spectral character signifies to a bat the presence of a moth. To develop a theory of

meaning, more must be known about how neurons code and transform sensory signals, about the neural basis of memory, learning and emotion and about the interaction of these capacities and the motor system. A neurally grounded theory of meaning may require revision of the very intuitions that now seem so secure and that are so freely exploited in Searle's arguments. Such revisions are common in the history of science.

Could science construct an artificial intelligence by exploiting what is known about the nervous system? We see no principled reason why not. Searle appears to agree, although he qualifies his claim by saying that "any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains." We close by addressing this claim. We presume that Searle is not claiming that a successful artificial mind must have *all* the causal powers of the brain, such as the power to smell bad when rotting, to harbor slow viruses such as kuru, to stain yellow with horseradish peroxidase and so forth. Requiring perfect parity would be like requiring that an artificial flying device lay eggs.

Presumably he means only to require of an artificial mind all of the causal powers relevant, as he says, to conscious intelligence. But which exactly are they? We are back to quarreling about what is and is not relevant. This is an entirely reasonable place for a disagreement, but it is an empirical matter, to be tried and tested. Because so little is known about what goes into the process of cognition and semantics, it is premature to be very confident about what features are essential. Searle hints at various points that every level, including the biochemical, must be represented in any machine that is a candidate for artificial intelligence. This claim is almost surely too strong. An artificial brain might use something other than biochemicals to achieve the same ends.

This possibility is illustrated by Carver A. Mead's research at the California Institute of Technology. Mead and his colleagues have used analog VLSI techniques to build an artificial retina and an artificial cochlea. (In animals the retina and cochlea are not mere transducers: both systems embody a complex processing network.) These are not mere simulations in a mini-computer of the kind that Searle derides; they are real information-processing units responding in real time to real light, in the case of the artificial retina, and to real sound, in the case

of the artificial cochlea. Their circuitry is based on the known anatomy and physiology of the cat retina and the barn owl cochlea, and their output is dramatically similar to the known output of the organs at issue.

These chips do not use any neurochemicals, so neurochemicals are clearly not necessary to achieve the evident results. Of course, the artificial retina cannot be said to see anything, because its output does not have an artificial thalamus or cortex to go to. Whether Mead's program could be sustained to build an entire artificial brain remains to be seen, but there is no evidence now that the absence of biochemicals renders it quixotic.

We, and Searle, reject the Turing test as a sufficient condition for conscious intelligence. At one level our reasons for doing so are similar: we agree that it is also very important how the input-output function is achieved; it is important that the right sorts of things be going on inside the artificial machine. At another level, our reasons are quite different. Searle bases his position on commonsense intuitions about the presence or absence of semantic content. We base ours on the specific behavioral failures of the classical SM machines and on the specific virtues of machines with a more brainlike architecture. These contrasts show that certain computational strategies have vast and decisive advantages over others where typical cognitive tasks are concerned, advantages that are empirically inescapable. Clearly, the brain is making systematic use of these computational advantages. But it need not be the only physical system capable of doing so. Artificial intelligence, in a nonbiological but massively parallel machine, remains a compelling and discernible prospect.

FURTHER READING

- COMPUTING MACHINERY AND INTELLIGENCE. Alan M. Turing in *Mind*, Vol. 59, pages 433-460; 1950.
- WHAT COMPUTERS CAN'T DO; A CRITIQUE OF ARTIFICIAL REASON. Hubert L. Dreyfus. Harper & Row, 1972.
- NEUROPHILOSOPHY: TOWARD A UNIFIED UNDERSTANDING OF THE MIND/BRAIN. Patricia Smith Churchland. The MIT Press, 1986.
- FAST THINKING in *The Intentional Stance*. Daniel Clement Dennett. The MIT Press, 1987.
- A NEUROCOMPUTATIONAL PERSPECTIVE: THE NATURE OF MIND AND THE STRUCTURE OF SCIENCE. Paul M. Churchland. The MIT Press, in press.

No other system
of keeping up
can compare with
**SCIENTIFIC
AMERICAN
Medicine.**



YOUR SYSTEM: a time-consuming, futile struggle to keep up with the information explosion.

The classic texts are convenient references—but the information they contain is obsolete before publication.

Like many physicians, you probably rely on the texts you first used in medical school. But even using the most recent editions, you find material that no longer reflects current clinical thinking—and they lack the latest information on such topics as herpes, oncogenes, AIDS, and photon imaging.

Reading stacks of journals alerts you to recent developments—but can't give you quick answers on patient management.

Struggling through the hundreds of journal pages published each month—even on only the really significant advances in the field—is arduous and memory-taxing. And it's a task that costs physicians valuable time—their most precious resource.

Review courses cover clinical advances—but, months later, do you recall the details of a new procedure or unfamiliar drug?

Seminars can also be costly and make you lose valuable time away from your practice—expenses that may amount to several thousand dollars. And, the speaker's skill often determines how much you learn.

**SCIENTIFIC
AMERICAN
MEDI**

OUR SYSTEM: a rewarding, efficient way to keep yourself up-to-date—and save hundreds of hours of your time for patient care.

A comprehensive, 2,300-page text in two loose-leaf volumes, incorporating the latest advances in medical practice as of the month you subscribe.

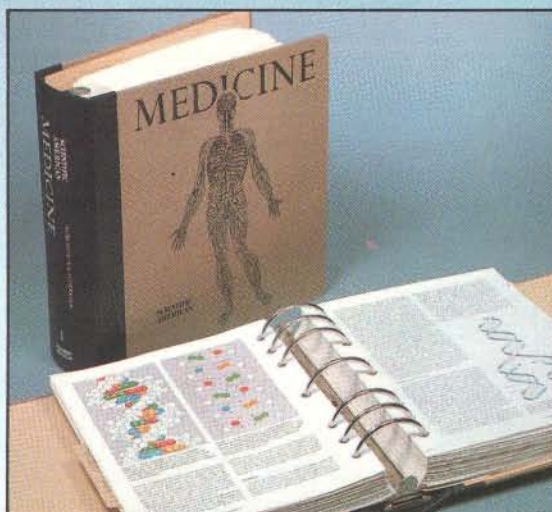
This superbly designed, heavily illustrated resource, called "the best written of all [the internal medicine] books" by JAMA (251:807, 1984), provides a practical, comprehensive description of patient care in 15 subspecialties. And, because the text is updated each month, the clinical recommendations reflect all the current findings. A practice-oriented index and bibliography of recent articles further enhance the efficiency of the text.

Each month, six to nine replacement chapters to update your text *plus* new references, an eight-page news bulletin, and a completely new index.

You'd have to read hundreds of journal pages each month—and memorize the contents—to get the same information SCIENTIFIC AMERICAN *Medicine* contains. With our updated text, you read only the information you really need. Our authors, largely from Harvard and Stanford, sort through the literature and monitor developments, incorporating the significant advances into our chapters.

At no additional cost, a 32-credit CME program, to save you valuable patient-care time and the expense of attending review courses.

Earn 32 Category 1 or prescribed credits per year with our convenient self-study patient management problems; each simulates a real-life clinical situation. Choose either the complimentary printed version or, at a modest extra charge, the disk version for your IBM® PC or PS/2™, Macintosh™, or Apple® (or compatible).



Your subscription includes:
the two-volume, 2,300-page loose-leaf text
and, each month, six to nine replacement chapters,
a newsletter, new references,
and a completely revised index.

Or call toll-free: 1-800-345-8112

CINE

SCIENTIFIC AMERICAN MEDICINE

415 MADISON AVENUE, NEW YORK, N.Y. 10017

- ☐ **Yes, I'd like to try the SCIENTIFIC AMERICAN *Medicine* system.** Please enter my subscription at a first-year price of US\$265* plus \$7 shipping for a total of US\$272.
- ☐ **Also enroll me in the CME program.** I'd prefer:
- ☐ the printed version at **no additional charge.**
 - ☐ the disk version at an additional \$US97.†
- Computer type or compatible:
- ☐ IBM® PC (5 1/4") 256K ☐ IBM® PS/2 (3 1/2")
 - ☐ Macintosh™ (512K)
 - ☐ Apple® IIc/IIe/IIgs
 - ☐ Apple® II+ with 80-col. card by: ☐ Apple® ☐ Videx®
- ☐ **Enroll me in the disk CME *only* at US\$152.†**
(Note model above.)
- ☐ Check enclosed* ☐ Bill me

Signature _____

☐ VISA ☐ MasterCard Exp. Date _____

Account No. _____

Name _____

Specialty _____

Address _____

City _____ State _____ Zip Code _____

*Add sales tax for IA, Ill, Mich., or N.Y. Allow 8 weeks for delivery. Add US\$10 for shipping to Canada. IBM is a registered trademark of International Business Machines Corporation. Apple is a registered trademark of Apple Computer, Inc. Videx is a registered trademark of Videx, Inc. Macintosh™ is a trademark of McIntosh Laboratory, Inc. and is used by Apple Computer, Inc. with its express permission.

†Please add sales tax for MI or NY.

Antisense RNA and DNA

Molecules that bind with specific messenger RNA's can selectively turn off genes. Eventually certain diseases may be treated with them; today antisense molecules are valuable research tools

by Harold M. Weintraub

It takes about 100,000 genes to make a human being. What exactly do they do, and how do they do it? To answer these questions, biologists must tinker with individual genes—in effect, remove or turn off the genes—and observe the effects on organisms or on individual cells. Studies of mutations have always afforded this information, but mutations are random by their nature, which has made systematic study of individual genes difficult (or, in the case of human beings and other complex and long-lived organisms, impossible).

With the recent advent of technology for cloning, or copying, genes during the past decade, it is now becoming realistic to think about selectively turning off or modifying the activity of any given gene. One method is—in principle—remarkably simple: create antisense RNA or DNA molecules that bind specifically with a targeted gene's RNA message, thereby interrupting the precise molecular choreography that expresses a gene as a protein. In this way viruses and bacteria regulate some genes during their life cycles. Today such an approach is practical enough for investigators to apply it to a broad range of problems.

The ability to deactivate specific genes holds great promise for medicine. For example, it may someday be possible to fight viral diseases with antisense RNA and DNA molecules

that seek and destroy viral gene products inside a person's cells. Such applications are in their infancy. In the meantime antisense technology is contributing to the birth of a new field, reverse genetics. Classical genetics usually studies the random mutations of all genes in an organism and selects the mutations responsible for specific characteristics; reverse genetics starts with a cloned gene of interest and manipulates it to elicit information about its function.

The traditional genetic approach relies on chemicals or radiation to delete or alter genes randomly. The treated organisms or cells and their progeny can then be observed, and mutated individuals that display characteristics of interest to the experimenter can be studied. The genetic approach has been tremendously successful with microorganisms, some plants and some invertebrate animals. Nevertheless, it is poorly suited to studies of vertebrates. Vertebrates have inconveniently long generation times; they often have small numbers of offspring, which limits how quickly interesting mutants can be produced; and their most intriguing mutations are usually lethal and consequently difficult to propagate and study.

Another shortcoming of any genetic approach is that the observable effect of a mutation may not reveal precisely what the mechanism of the mutation is. For example, a mutation may first be detected as a microorganism's diminished ability to grow on a source of sugar. Is the mutation altering an enzyme that digests the sugar, or is it blocking the cell's uptake of the sugar? Is it perhaps activating enzymes that cause the sugar to be stored instead of digested? Genetics alone cannot provide the answer; genetics can identify the range of genes that influence a process, but additional approaches are often needed to probe exactly what a particular gene does. The new approach involving antisense

technology is one attempt to overcome some of these problems.

To understand how antisense technology works, it is first necessary to review the fundamentals of gene structure and expression. A gene is a coded blueprint for a protein; the code is written in the precisely ordered sequences of four nucleotide bases—adenine (A), thymine (T), guanine (G) and cytosine (C)—that make up molecules of DNA. In addition to the strong linking bonds that these bases can make within DNA strands, A's can form weak bonds with T's in other strands, and G's can form similar bonds with C's.

For this reason, DNA in organisms usually exists as a double helix, or duplex, consisting of two coiled DNA strands. In each duplex, a base on one strand is bound to its complementary base on the other strand; for example, a "sense" sequence that reads A-T-G-C-T-C on one strand pairs with an "antisense" sequence, T-A-C-G-A-G, on the other strand. The high specificity of base pairing is important during DNA replication, when the two complementary strands of each duplex separate: each strand serves as a template for reconstructing its partner, and the result is two identical duplexes.

Base-pairing specificity is also important when genetic information is decoded to make proteins. DNA does not make proteins directly; instead an intermediary molecule of RNA is created for the job. RNA is made of the same bases as DNA except that the base T is replaced by the base uracil (U), which can also pair with A. During transcription, the first stage of reading the genetic code, the sense strand of a gene separates from its antisense partner. Enzymes then assemble an RNA molecule that complements the sequence on the antisense DNA strand. This messenger RNA eventually migrates to cell structures called ribosomes, which read the encoded in-

HAROLD M. WEINTRAUB is a member of the division of basic sciences at the Fred Hutchinson Cancer Research Center in Seattle. He received his Ph.D. from the University of Pennsylvania School of Medicine in 1971 and his M.D. in 1973. After completing his postdoctoral work at the Medical Research Council in Cambridge, England, he spent four years at Princeton University in the department of biochemical sciences. His current research interests are gene regulation and development.

formation and string together the appropriate amino acids to form the encoded proteins.

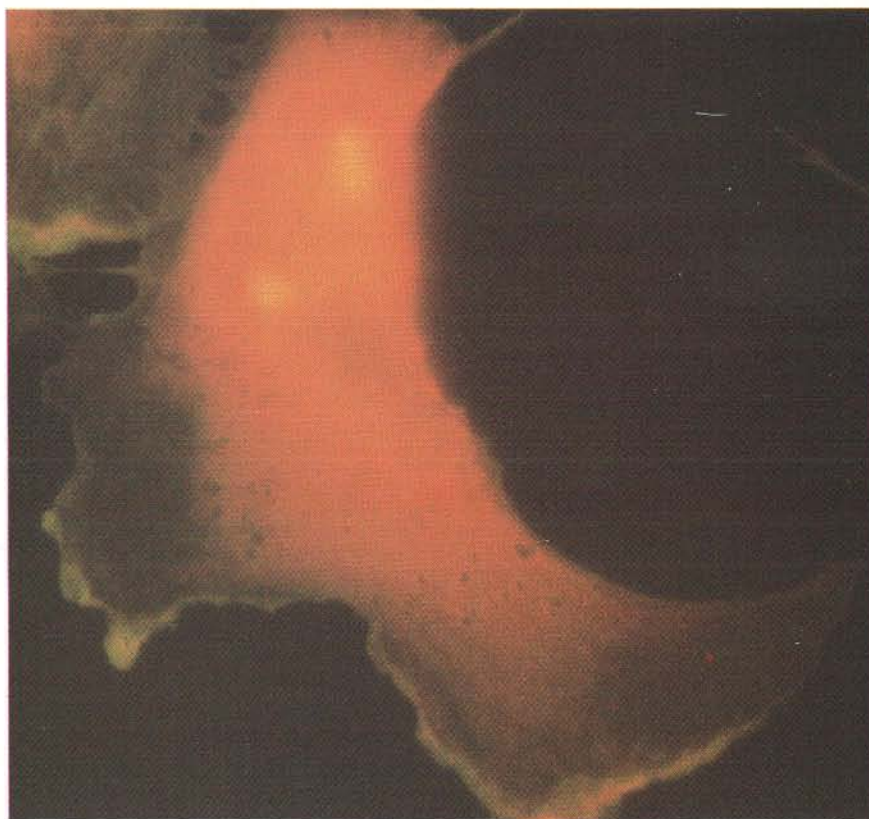
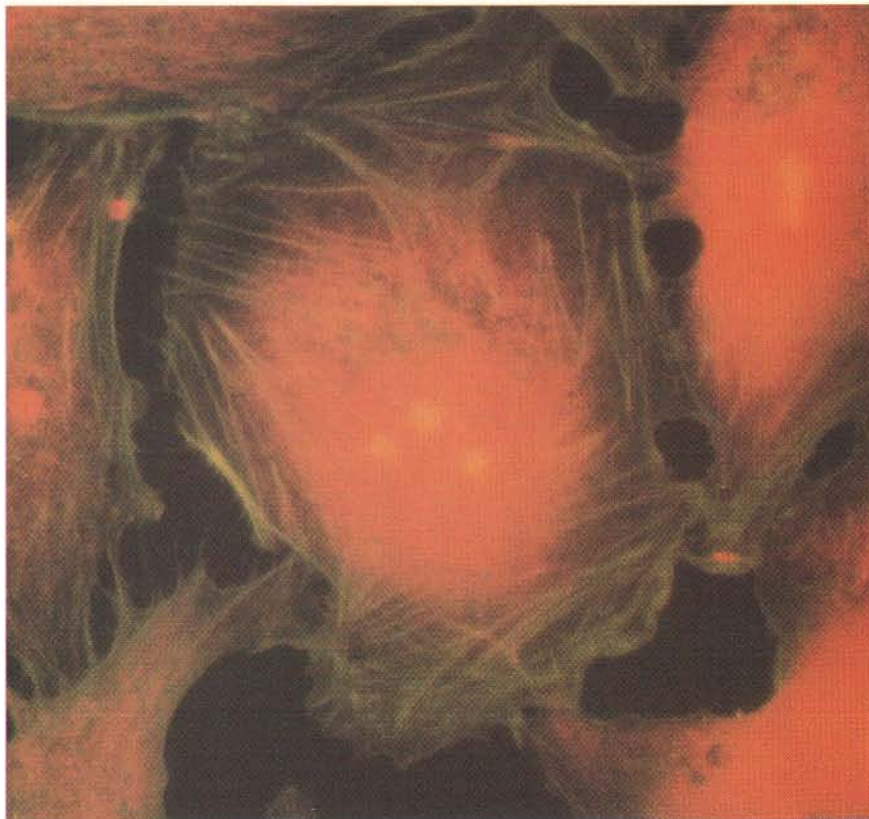
The antisense strand of DNA in the genes is the template for messenger RNA; the messenger RNA carries the structural code for a protein. But what of the sense DNA strand? Does it produce antisense RNA, and if so, what does this RNA do?

An important discovery about the natural biological function of antisense RNA molecules was made in 1981 by Jun-ichi Tomizawa at the National Institutes of Health (NIH). Tomizawa was studying the replication of a plasmid, or small double-strand ring of bacterial DNA, called ColE1. In this plasmid, DNA replication begins at a specific sequence called the origin. First the primer, a short chain of DNA, opens up the DNA double helix and hybridizes, or pairs, with the origin. The enzyme DNA polymerase then adds A's, T's, C's and G's to the RNA primer, constructing a new DNA strand that is complementary to the origin-containing strand. The number of copies made of the plasmid genetic material depends on the number of available RNA primers inside the cell.

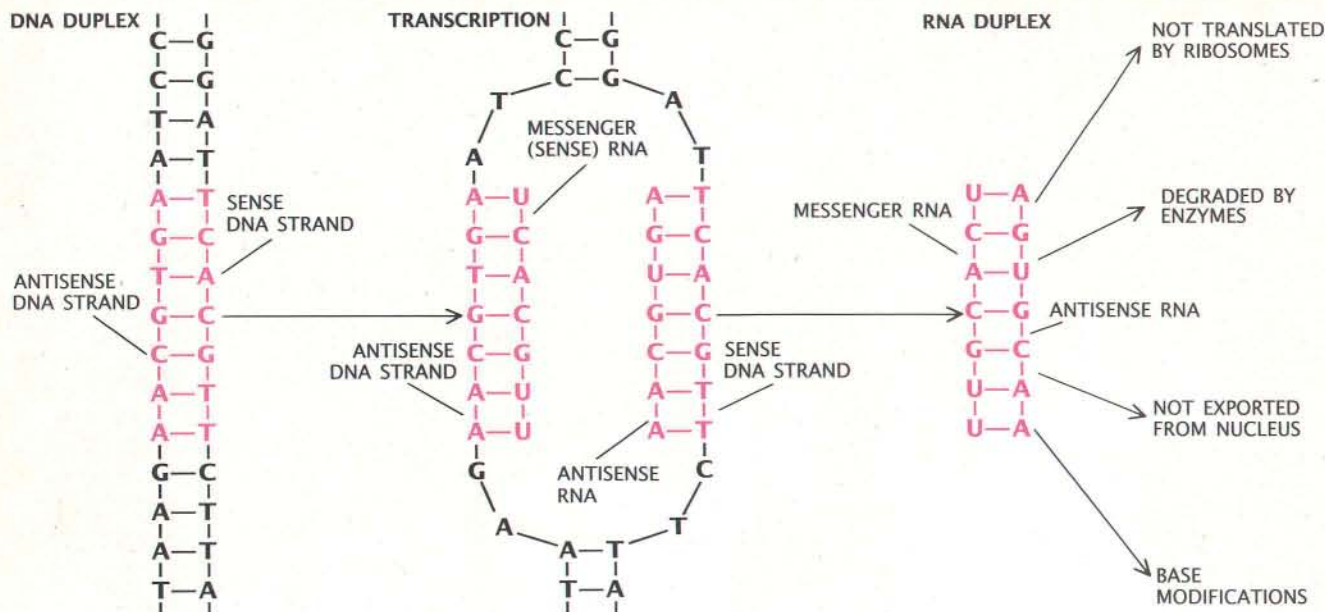
Tomizawa discovered that the availability of RNA primers is controlled not by their total concentration but rather by the ratio of primers to specific inhibitor molecules. He went on to show that these inhibitors are RNA molecules transcribed from the DNA strand complementary to the one that produces the RNA primer. In other words, the inhibitors are the antisense products of the sense DNA strands.

Just as the sense and antisense DNA strands are complementary, so too are the sense RNA primers and the antisense inhibitor molecules. Consequently, the sense RNA and the antisense RNA can hybridize with one another. In this duplexed state, the RNA primer cannot initiate DNA replication, because it cannot pair with the origin of the plasmid.

The function of antisense RNA goes beyond regulation of DNA replication; it also extends to regulation of transcription. In 1983 Nancy E. Kleckner of Harvard University conducted a series of elegant experiments that described how antisense RNA in bacteria controls the synthesis of the enzyme transposase. During transcription, messenger RNA encoding the enzyme is produced from the transposase gene. When the bacteria also transcribe antisense RNA from the sense strand of the transposase gene, the antisense



ANTISENSE RNA molecules can selectively inhibit the activity of genes and block the production of specific proteins in living cells. For example, normal cells in culture (*top*) make a network of thin structural filaments (*green*) from the protein actin; these actin filaments help give the cells a smooth, round shape. Cells that have been injected with antisense-actin RNA (*bottom*) lose much of their actin framework and become flatter. Antisense techniques can probe the functions of individual genes.



BASE-PAIR SPECIFICITY is the key to antisense RNA's inhibition of genes. In the DNA duplex of a gene, there are weak bonds between opposing pairs of adenine (A) and thymine (T) bases and between guanine (G) and cytosine (C) bases. The matched sense and antisense strands of DNA complement each other. During the transcription of DNA into RNA, the antisense DNA strand acts as a template for assembling a complementary (sense) messenger RNA molecule (in which U—for uracil—substitutes for T). A single-strand messenger RNA is translated into a protein on the cellular organelles called ribosomes.

Messenger RNA is the only transcription product of most genes; however, some genes are regulated by the additional transcription of an antisense RNA from the sense DNA strand. If an antisense RNA is made, then the antisense RNA and the messenger RNA will bind with each other. A variety of factors may then prevent the translation of protein: the RNA duplex may be rejected by the ribosomes; it may be degraded by enzymes; it may never leave the nucleus; or the A bases may be modified chemically to become different inosine bases, thereby scrambling the genetic code on the messenger RNA.

RNA binds specifically with the sense messenger RNA and prevents the ribosome from translating the encoded information into a protein.

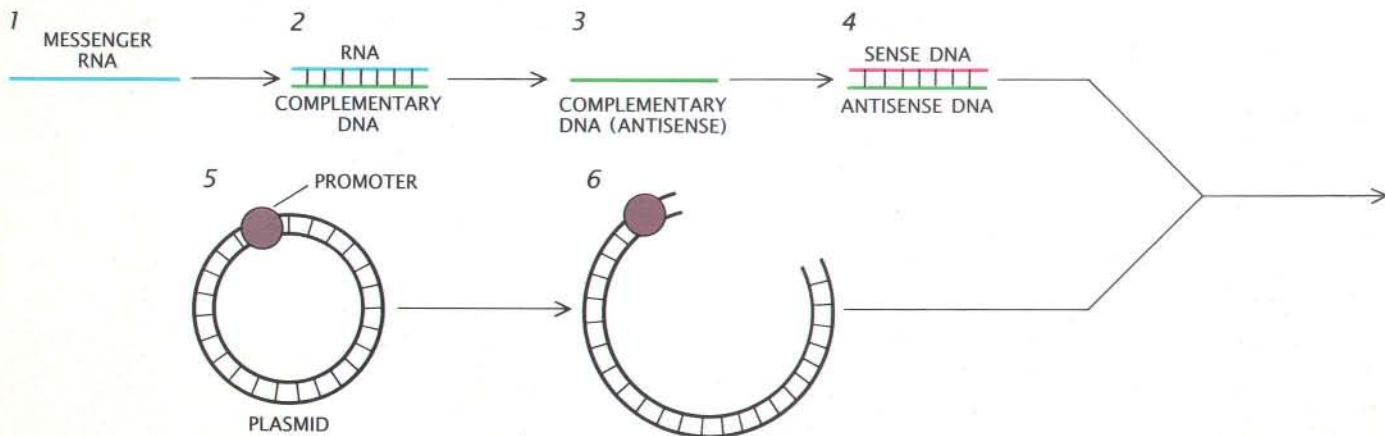
Investigators have shown that regulation or inhibition of gene activity with antisense RNA seems to be universal among viruses and bacteria. The mechanism controls many stages of cell metabolism. There is some tentative evidence that antisense RNA

may also have a natural role in more complex cells, but this has not yet been proved.

In 1983 the initial work of Tomizawa suggested to me and to Jonathan G. Izant, who is now at Yale University, that we could direct antisense RNA against any cloned gene and perhaps inhibit the translation of its messenger RNA. We thought that this approach might permit us to inactivate

specific genes much as mutations could, but with higher selectivity. One of the first questions confronting us was how to manufacture the antisense RNA. We decided to exploit recombinant-DNA technology to create artificial genetic elements, called expression vectors, that would make antisense RNA when they were inserted into cells.

For a test gene we chose the one for



EXPRESSION VECTORS that make antisense RNA can be engineered from DNA duplexes in the laboratory and introduced into cells. An isolated messenger RNA molecule (1) can serve as a template for making a complementary strand of antisense

DNA (2). This DNA strand (3) can, in turn, act as a template for making the sense DNA strand and creating a DNA duplex (4). A plasmid, or double-strand ring of bacterial DNA (5), can be cut by restriction enzymes near a promoter region (6). The DNA

the enzyme thymidine kinase (TK) from herpes simplex virus (HSV). This enzyme converts thymidine, a molecular precursor for the T base in DNA, into a form that can be added to growing strands of DNA. The cells with which we chose to work were from a standard mouse cell-culture line that had a mutated TK gene and could incorporate thymidine into replicating DNA. If expression vectors containing the HSV-TK gene were injected into these mouse cells, the cells began to incorporate thymidine normally.

To construct the expression vectors for antisense-TK RNA, we began with the cloned gene for HSV-TK itself. With so-called restriction enzymes that cleave DNA at specific base sequences, we snipped the HSV-TK gene out of double-strand expression vectors and permitted the genes to reinsert themselves into the vectors at random. Half of the TK genes returned to the vectors in their original orientation. The others, however, were reversed, so that the sense-TK sequence was inserted into the strand that had originally contained the antisense-TK sequence and vice versa. Transcription of these expression vectors, we believed, would produce antisense-TK RNA instead of messenger-TK RNA.

When we injected these antisense expression vectors into cells that had previously received HSV-TK expression vectors, the ability of the cells to incorporate thymidine did indeed diminish significantly. As we had suspected, the presence of the antisense expression vector was inhibiting the activity of the sense expression vector. We were also able to show that antisense-TK RNA from chickens could inhibit the activity of chicken TK genes.

Yet antisense-TK RNA from a chicken failed to inhibit TK genes from a virus and vice versa: because chicken TK genes and viral TK genes are dissimilar, their RNA products do not hybridize. These results showed that antisense inhibition can be highly specific.

Through these experiments we were able to demonstrate that specific antisense RNA's could inhibit the function of cloned target genes. Working with bacteria, Masayori Inouye and his colleagues at the State University of New York at Stony Brook and Sidney Pestka and his colleagues at the Roche Institute of Molecular Biology in Nutley, N.J., had similar successes with expression vectors for antisense RNA.

It was clear that antisense RNA could affect cloned target genes inserted into cells; however, for antisense technology to be helpful in the analysis of vertebrate-gene functions, we also needed to prove that antisense RNA could turn off an endogenous cellular gene. For this reason, we next chose to work on the gene for actin, a protein that is a major component of a structural "cytoskeleton" that helps cells to move and maintain their shape. When we injected expression vectors for antisense-actin RNA into cells, the cells developed defective cytoskeletons and appeared flatter. Once again the antisense RNA was inactivating its target gene.

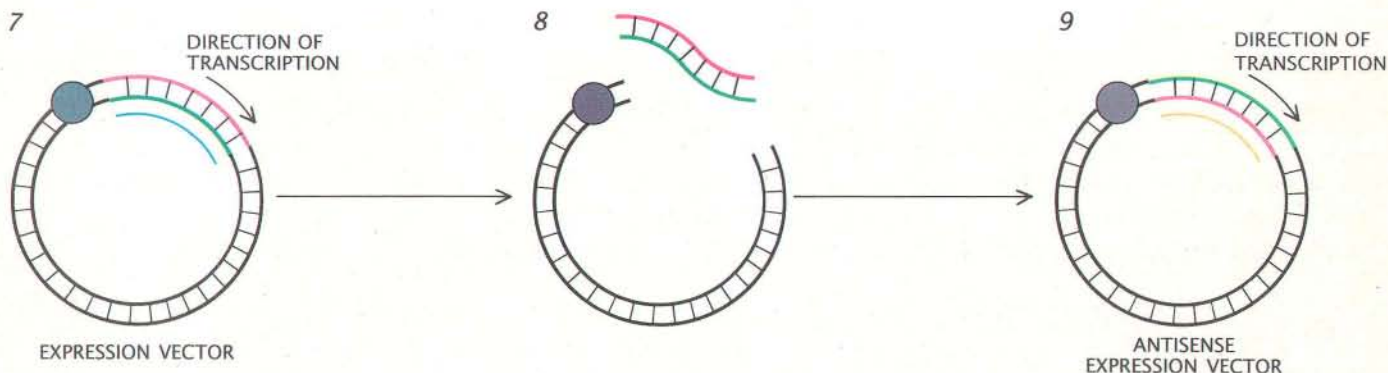
Because actin is an essential protein, cells that do not express actin will die. This problem can be circumvented with inducible promoters. A promoter is a part of a gene that controls the initiation of transcription for the gene product. An inducible promoter

is one that initiates transcription only in the presence of a certain inducing stimulus. Promoters inducible by specific ions, by elevated temperatures ("heat shock") and by various hormones have been well studied. Having added an inducible promoter to an antisense expression vector, an experimenter can start or stop the production of antisense RNA (and of the gene product it affects) at any time, in any individual cell, by supplying or removing the inducer.

My colleagues and I have added inducible promoters that respond to steroid hormones to the expression vectors for antisense-TK RNA. We have demonstrated that under conditions in which cells need active TK genes to grow, steroids inhibit the growth of cells containing these expression vectors. These cells presumably lack TK and cannot synthesize DNA. Susan Lindquist of the University of Chicago has had similar results with antisense expression vectors controlled by heat-shock inducible promoters.

It may be possible to modify antisense expression vectors so that the resulting antisense RNA has specific functional properties. Thomas R. Cech of the University of Colorado at Boulder has shown that certain naturally occurring RNA sequences can cleave RNA targets with which they hybridize [see "RNA as an Enzyme," by Thomas R. Cech; *SCIENTIFIC AMERICAN*, November, 1986]. These RNA sequences, called ribozymes, cleave at a point where there is a specific triplet of nucleotide bases, G-U-C, in the target. It is statistically likely that a G-U-C triplet will appear by chance at least once in most messenger RNA's.

Once the location of the triplet in



duplex can be spliced into the plasmid to create an expression vector (7). The effectiveness of this expression vector can be tested in cells: when the promoter initiates transcription, the expression vector will make copies of the original messenger

RNA (blue). If the added DNA is then cut out of the expression vector with restriction enzymes (8), it can reinsert itself into the ring with the opposite orientation (9). During transcription, this expression vector will make antisense RNA (yellow).

a specific messenger RNA has been determined, investigators can insert DNA for a ribozyme into the corresponding position in the antisense expression vector. The vector will then produce antisense RNA containing a ribozyme. When this ribozymal antisense RNA hybridizes with its messenger RNA target, it cleaves the messenger molecule. This process has been demonstrated in the test tube; it also appears to take place in certain plants infected with viruses that produce ribozymelike RNA's. Whether this approach can be successfully adapted for investigations of gene functions remains to be explored.

Antisense-RNA methods for inhibiting gene activity need not rely on expression vectors. Strands of antisense RNA can be synthesized in the laboratory and injected directly into cells. In 1985 Douglas A. Melton of Harvard University and Richard M. Harland in my laboratory independently demonstrated that injections of antisense RNA into the large egg-producing cells (oocytes) of frogs could inhibit the translation of a corresponding sense RNA injected previously. Injecting antisense-actin RNA into cells inhibits the formation of a cytoskeleton much as injections of antisense expression vectors do.

Producing large amounts of antisense RNA for experiments has been greatly simplified by a clever technique designed in 1985 by Barbara J. Wold and Stuart K. Kim of the California Institute of Technology. Their technique takes advantage of the fact that cultured cells occasionally "amplify" genes by making multiple copies of them. Wold and Kim linked an antisense expression vector to a gene for an essential enzyme and inserted the vector into cells. With an inhibitor of the enzyme, they then selected for cells in subsequent generations that had amplified the enzyme gene and the expression vector. These cells

manufactured large quantities of antisense RNA.

Antisense RNA's are not the only antisense molecules that can latch onto messenger RNA's and prevent the translation of protein. Short complementary strands of DNA can also hybridize with messenger RNA's. Antisense oligonucleotides, strands of DNA only 15 to 25 bases long, can be created in the laboratory and introduced into cells, where they will have inhibitory effects like those of antisense RNA. This notion was first described by Paul C. Zamecnik of the Worcester Foundation for Experimental Biology in Massachusetts, who used antisense DNA oligonucleotides in an attempt to inhibit Rous sarcoma virus (RSV) from transforming cultured chicken cells into a cancerous state. It appears that some cells can absorb enough antisense oligonucleotides from their environment for the oligonucleotides to bind with messenger RNA's from viral genes.

The specificity of an antisense oligonucleotide depends on its length. The longer an oligonucleotide is, the more likely it is to bind to one and only one DNA or RNA target. Specificity is essential because it would be disastrous if an oligonucleotide inhibited the wrong gene. For example, in the chromosomes of a human cell, there are three billion pairs of nucleotide bases. If all four bases are present in roughly equal numbers and are distributed randomly throughout the genes, then, on average, no 12-base sequence should appear more than once.

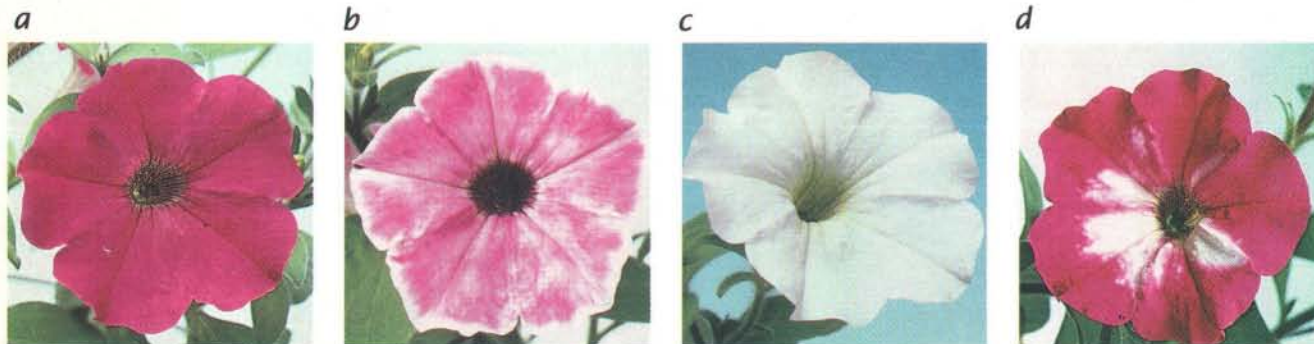
The functional chemistry of oligonucleotides can be tailored in the laboratory to meet the needs of experimentalists. Paul O. P. Ts'o and Paul S. Miller of Johns Hopkins University have pioneered this approach. They made oligonucleotides with uncharged sugar-phosphate "backbones"; such oligonucleotides can enter a cell through its outer membrane

more easily than can ordinary oligonucleotides. Phosphorothioate oligonucleotides, in which sulfur atoms substitute for oxygen atoms in the sugar-phosphate backbone, have also been synthesized; these molecules are less vulnerable to nucleases (cellular enzymes that degrade RNA and DNA) and can therefore remain at high, effective concentrations inside cells for longer periods.

Peter B. Dervan of the California Institute of Technology has extended this approach even further. He has attached a reactive chemical group, iron-linked ethylenediaminetetraacetic acid (EDTA-Fe), to an antisense oligonucleotide. If this oligonucleotide were to hybridize with a messenger RNA, the reactive group would inactivate the RNA by cleaving it at the site of hybridization.

Claude Hélène and his colleagues at the National Museum of Natural History in Paris have used an acridine attachment to an oligonucleotide; acridine is a planar molecule that can insert itself between the base pairs in a DNA double helix. Hélène has found that the presence of acridine on the oligonucleotide increases the energy with which the antisense molecule binds to the messenger RNA target and also increases the rate at which cells absorb the oligonucleotide from their environment. He and J.-J. Toulmé have recently taken this approach in cell cultures to kill *Trypanosoma brucei*, an African microorganism that causes the parasitic disease trypanosomiasis: an antisense oligonucleotide was directed against a sequence found on most, if not all, trypanosome messenger RNA's.

Hélène's success at killing trypanosomes with antisense oligonucleotides illustrates the great potential of these molecules as therapeutic agents. Needless to say, much effort is now being directed toward developing safe and effective antisense oligonucleotides for medical purposes. Such



PIGMENT GENES in petunias can be inhibited by antisense molecules. A normal flower (a) shows solid coloration. Three

other flowers (b-d) contain antisense expression vectors that block the enzymatic production of red pigment in some cells.

agents could be especially important in the treatment of viral diseases, because investigators often know the exact nucleic acid sequences present in disease-causing viruses. In tissue culture, antisense oligonucleotides have inhibited infections by herpesviruses, influenza viruses and the human immunodeficiency virus that causes AIDS.

It may also be possible to target antisense oligonucleotides against dangerously mutated oncogenes, the genes that transform normal cells into cancer cells. The major challenge is to make antisense agents that will inactivate the mutated oncogene but not their normal precursors, or proto-oncogenes, which are generally essential for cell survival.

Such therapeutic applications for antisense technology are still on the horizon. But antisense methods are already making contributions to the understanding of gene function, particularly in the areas of development, cell growth and cell division. One noteworthy example is the work that has been done on the *src* gene, a cellular growth-promoting gene that is present in a mutated form in some viruses. The *src* gene encodes a protein that controls the activity of other proteins by modifying them chemically. It is an abnormal form of the *src* gene in the Rous sarcoma virus that transforms infected chicken cells.

In 1987 Paul E. Neiman and his colleagues at the Fred Hutchinson Cancer Research Center in Seattle showed that antisense methods can prevent RSV from transforming cells. They developed benign viruses that carried expression vectors for antisense RNA against the RSV *src* gene. When cells were infected with these special viruses and then exposed to RSV, the cells did not transform; in a sense, they had been "immunized" against RSV.

Understanding of the function of *src* has been further elaborated by David I. Shalloway of the Pennsylvania State University and Joseph B. Bolen of NIH and their colleagues. They conducted experiments on the polyoma virus, which can transform cells much as RSV does. The transforming gene in polyoma virus is not *src* but one for a protein called middle *T* antigen. Nevertheless, if expression vectors encoding antisense-*src* RNA are introduced into polyoma-transformed cells, the cells lose their cancerous characteristics and revert to their normal form. This effect suggests that the polyoma virus transforms cells by activating the cellular *src* gene product. It fol-

lows, then, that *src* can become an oncogene both directly (through mutation) and indirectly (by having polyoma middle *T* antigen bind to it).

Other genes regulating growth and development have also been studied with antisense techniques. In 1988 Helmut Ponta, Peter Herrlich and their colleagues at the University of Karlsruhe probed the interactions of three oncogene proteins: the *fos* gene product (which helps with the transcription of certain genes in the nucleus), the *ras* gene product (which brings chemical signals into the cell across the outer membrane) and the *sis* gene product (a membrane-bound receptor molecule affecting growth). Each of these three gene products can transform cells independently. Antisense RNA against the *fos* gene turned out to block the malignant effects of not only *fos* but also the abnormal *ras* and *sis* genes; it is therefore likely that *fos* mediates the activities of these genes.

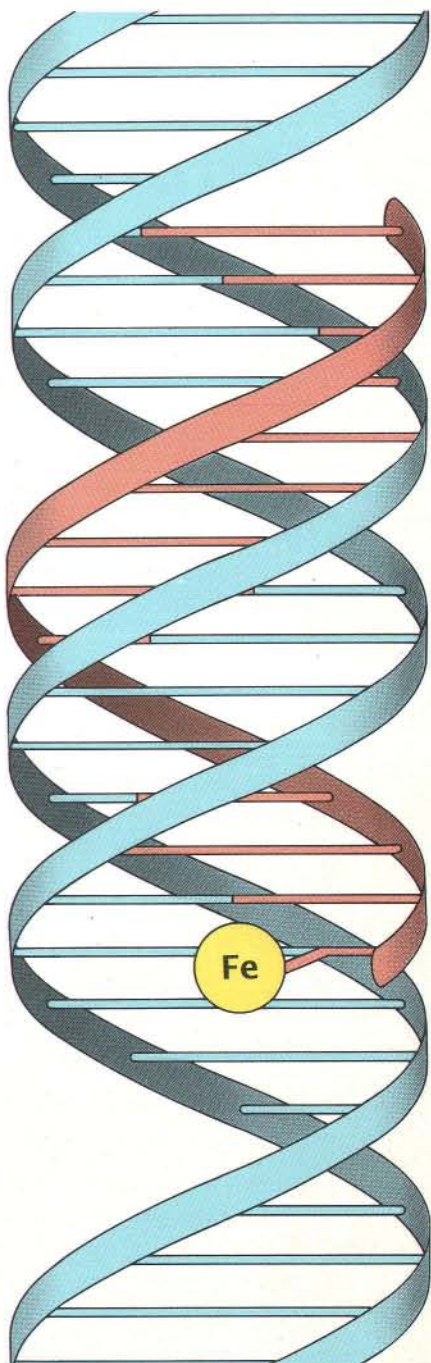
The cycle of cell growth and reproduction, it has long been suspected, is determined by a specific protein that gradually accumulates during cell growth and is destroyed rapidly during cell division. Such a protein, appropriately called cyclin, was identified in 1983 by Joan V. Ruderman of Harvard Medical School and Tim Hunt of the University of Cambridge. In 1989 Hunt and his colleagues showed that antisense DNA oligonucleotides directed against cyclin-messenger RNA arrest cell growth and division, which suggested that cyclin is important to the cycle of growth and reproduction.

One of the most important applications for antisense technology is the creation of particularly revealing mutations on demand. Herbert Jäckle and his colleagues at the University of Munich were the first to mimic mutations in *Drosophila* fruit-fly embryos with antisense technology. By inactivating specific genes with antisense RNA, they were able to produce fruit flies whose characteristics matched those of flies known to have mutations in those genes.

Such phenocopying has also been achieved in more complex organisms, including mice. "Shiverer" is the name given to a type of mutant mouse that has a defective nervous-system protein called myelin basic protein. Because of this defect, the myelin sheaths surrounding and insulating the nerve fibers are imperfect, and the shiverer mice shake uncontrollably. When antisense expression vectors against myelin basic protein are inserted into normal mouse embryos,

the resulting mice tremble like shiverer mutants. Antisense methods may provide mouse models for the study of human disease states.

Antisense techniques are not limited to phenocopying mutations that already exist. Entirely new forms can



TRIPLE HELICAL DNA is the result of a third DNA strand binding with a specific target region on a normal DNA duplex. In this triple-strand condition, the gene encoded by the duplex may not be transcribed. If an iron-containing molecule (Fe) is chemically bound to the end of the third DNA strand, it will cleave the DNA duplex and destroy the associated gene.

also be created to suit the purposes of investigators. Antisense expression vectors have been injected into petunias to inhibit an enzyme that produces pigment in the flowers. The resulting flowers display unusual pigment patterns. In tomatoes, antisense expression vectors can inhibit an enzyme that breaks down the tough, fibrous walls around plant cells; treated tomatoes ripen more slowly than untreated ones, which suggests that it may be possible to create tomatoes that can be transported more easily over great distances without spoiling or bruising. Antisense RNA directed against plant viruses has also been used to breed disease-resistant species of tobacco.

Many important refinements of antisense technology are still needed, and many important questions must still be answered. For example, it should be possible to modify antisense oligonucleotides chemically so that they can be introduced into cells more efficiently or be bound with their targets more effectively or can modify their targets. In addition, certain parts of messenger RNA's may be more susceptible than other parts to inhibition by antisense RNA. As these susceptibilities become better understood, it should be possible to design more effective antisense RNA molecules. Methods must also be designed for ensuring that antisense RNA's and DNA oligonucleotides are maintained at a high level in cells. This could be accomplished by linking an antisense expression vector to a highly active promoter region or by modifying the antisense oligonucleotide chemically so that it is less vulnerable to degradation by nucleases.

A greater understanding of the precise mechanisms whereby antisense RNA inhibits the production of proteins is also essential. Research suggests that antisense RNA can act both within the cell's nucleus and in the cytoplasm and that it may arrest protein translation by doing more than hybridizing with messenger RNA's. Antisense RNA may act early within the nucleus to prevent messenger RNA's from being spliced and modified in essential ways. Observations by Wold and Kim suggest that antisense RNA can prevent the export of messenger molecules from the nucleus to the ribosomes in the cytoplasm. It also appears that, at least in some cells, duplexed RNA is a sensitive target for potent enzymes. Recent experiments involving antisense RNA in mouse embryos have shown that cellular nucle-

ases cut messenger RNA's at points where antisense RNA binds to them.

A surprising observation was made by Brenda Bass in my laboratory. She discovered that when duplexes of antisense and messenger RNA are injected into cells, some cellular activity separates the duplex into two strands. Further investigation revealed that the adenines in the A-U base pairs had been modified by the cell to become inosines (I's), another type of nucleotide base. These I-U base pairs are less stable, causing the duplex to fall apart. More important, the substitution of inosines for adenines in the messenger RNA should change the messenger's coded message, thereby preventing it from producing its original protein product. This substitution effect may enhance the inhibitory effects of antisense RNA in some cell types.

Evidence that A-to-I conversions occur was found recently by Marc W. Kirschner and David Kimelman of the University of California at San Francisco. They observed that a natural messenger RNA for a growth-and-differentiation factor is associated in embryos with an antisense partner and that the A's in these RNA's are converted to I's. Martin Billeter of the University of Zurich and his colleagues have observed similar A-to-I conversions in viruses that infect human beings and that probably go through a double-strand RNA form during their life cycles.

A completely new way of exploiting the exquisite specificity of base sequences to inactivate specific genes has recently been developed. It goes back to important observations made three decades ago by Alexander Rich, David R. Davies and Gary Felsenfeld of NIH and extended by Jacques R. Fresco of Princeton University. They demonstrated that DNA can occasionally form triple helices instead of duplexes: an extra strand associates specifically with certain sequences in the paired strands. Bases in the third strand recognize and bind to specific base pairs in the duplex.

In 1987 Dervan took advantage of this fact to form a triple helix whose third strand was a DNA oligonucleotide directed against a sequence in a normal duplex DNA. Attachment of a third strand may block the expression of the gene in the original duplex, possibly by blocking the attachment of control proteins or enzymes essential to transcription. Dervan has also outfitted the oligonucleotide with the cleaving agent EDTA-Fe. The modified oligonucleotide cuts the DNA duplex at the site of triple-helix formation. As

yet the mechanism whereby a third strand binds to a DNA duplex is only partially understood, and the code by which a third base "recognizes" a specific base pair is not completely worked out. It should be possible, however, to design oligonucleotides that will predictably bind with and destroy specific double-strand DNA targets such as viral genes. This is an active area of current research.

Another highly promising technique is targeted homologous recombination: cloned DNA containing a mutated copy of a target gene can be introduced into cells; this DNA somehow finds its target in the cell's nucleus and replaces it in its normal chromosome context. Leland H. Hartwell of the University of Washington and his colleagues have shown that cloned genes can be engineered to overexpress themselves in cells, thereby resulting in phenotypes that provide clues to the functions of the genes. Cloned genes can be programmed to express themselves inappropriately in cells; my colleagues and I have shown that the expression of a muscle-determining gene, *MyoD*, in a fat cell will convert the cell to muscle.

The challenge to think of new ways to manipulate the activity of cellular genes is spurring the development of antisense technology and these other techniques. No doubt with more information, more sophisticated technology and additional imagination, new approaches will emerge that will complement those already in use. The new field of reverse genetics is rapidly providing inroads into the understanding of gene function; with luck it will eventually enhance medicine's ability to understand and treat disease.

FURTHER READING

- ANTI-SENSE RNA AS A MOLECULAR TOOL FOR GENETIC ANALYSIS. Harold Weintraub, Jonathan G. Izant and Richard M. Harland in *Trends in Genetics*, Vol. 1, No. 1, pages 23-25; January, 1985.
- THE ROLE OF ANTISENSE RNA IN GENE REGULATION. Pamela J. Green, Ophry Pines and Masayori Inouye in *Annual Review of Biochemistry*, Vol. 55, pages 569-597; 1986.
- ANTISENSE OLIGODEOXYRIBONUCLEOTIDES: AN ALTERNATIVE TO ANTISENSE RNA FOR ARTIFICIAL REGULATION OF GENE EXPRESSION—A REVIEW. J.-J. Toulmé and C. Hélène in *Gene*, Vol. 72, No. 1, pages 51-58; December 29, 1988.
- MODULATION OF EUKARYOTIC GENE EXPRESSION BY COMPLEMENTARY RNA OR DNA SEQUENCES. Alexander R. van der Krol, Joseph N. M. Mol and Antoine R. Stuitje in *BioTechniques*, Vol. 6, No. 10, pages 958-975; 1988.

Introducing...

CARE of the SURGICAL PATIENT

from SCIENTIFIC AMERICAN Medicine

Because the quality of your care depends on the quality of your information.

Treating pre and post operative patients poses a unique set of challenges. Yet in one way it's no different than any other practice issue.

Doing it well takes the right information.

That's why SCIENTIFIC AMERICAN Medicine is pleased to announce the publication of CARE of the SURGICAL PATIENT.

The definitive resource on pre and post-operative care.

CARE of the SURGICAL PATIENT gives you ready access to the most authoritative and current information on pre and post-operative standards available anywhere.

Written and designed by prominent surgeons under the supervision of the American College of Surgeons' Committee on Pre and Postoperative Care, CARE of the SURGICAL PATIENT

provides two volumes – over 1,500 pages – of practical information on both critical and elective care.

And, CARE of the SURGICAL PATIENT is updated twice a year, with each surgeon-author reviewing his own specialty. Updates include new information on significant topics, such as current developments on AIDS.

In short, CARE of the SURGICAL PATIENT presents the standards for pre and postoperative treatment. You simply won't find a more important resource. Or one that organizes its information in such an intelligent way.

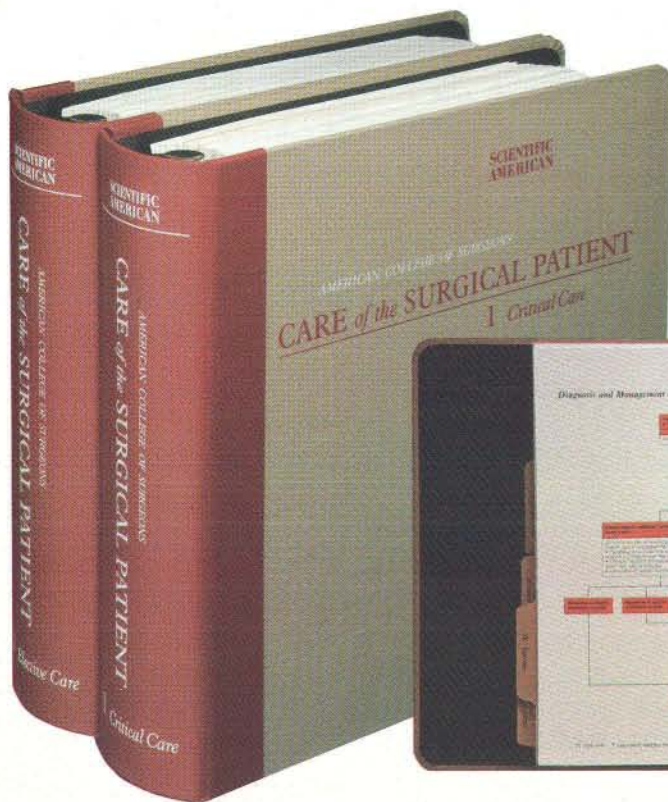
A unique system for rapid information retrieval.

CARE of the SURGICAL PATIENT is designed to get you the information you need, the way you need it.

Quickly. And intelligently.

The key is the system's three-part format. Chapters begin with a full page algorithm – the relevant facts at a moment's glance. Next, there's a detailed explanation of each element laid out in the treatment pathway. The third section covers etiology, pathobiology, and relevant clinical advances, as well as current references.

You choose the level of detail you need at the moment. Without having to wade through everything else. And unlike most texts, CARE of the SURGICAL PATIENT covers topics in order of urgency, instead of by organ system. Which means you have access to information as it relates to the real world treatment of the patient.



Try CARE of the SURGICAL PATIENT Free for 30 days.

You'll find it the most valuable resource on pre and post-operative care that's ever been published. And if you're not satisfied, just return it. No risk. No obligation.

CARE of the SURGICAL PATIENT, from SCIENTIFIC AMERICAN Medicine. No other resource helps you keep up better. And the better you keep up, the better your care.

☐ **YES, please send me CARE of the SURGICAL PATIENT.**

I will receive the two-volume, 1,500 page set and one update, at a first-year price of US\$200. (Sales tax added for MI or NY.) If not completely satisfied, I may return the books within 30 days for a full refund.

☐ Check enclosed ☐ MasterCard ☐ VISA ☐ Bill me

Acct.# _____ Exp. Date _____

Name _____

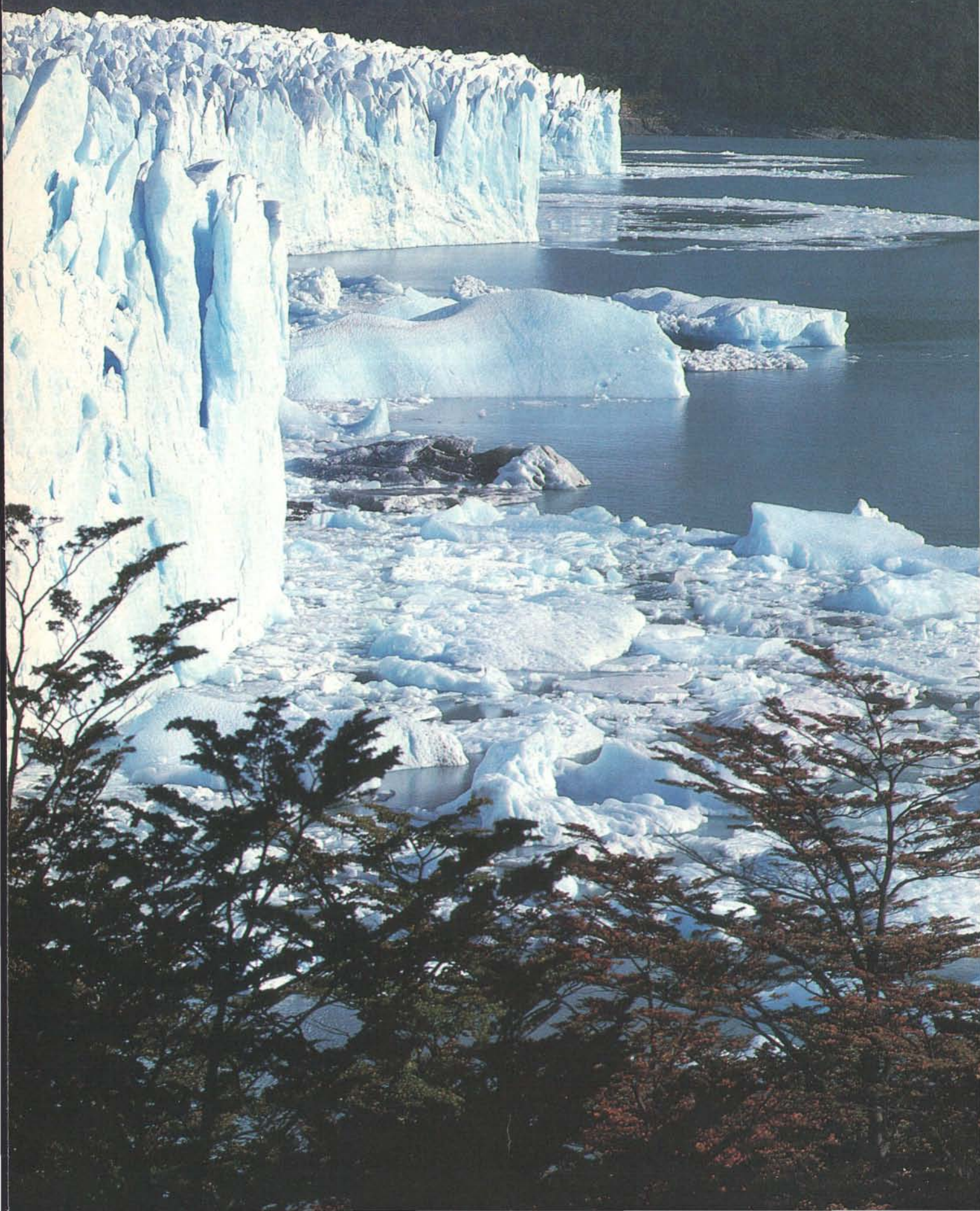
Address _____

City _____ State _____ Zip _____

Specialty _____

Or call toll free 1-800 345 8112.

SCIENTIFIC AMERICAN MEDICINE 415 Madison Avenue,
New York, NY 10017 9510



What Drives Glacial Cycles?

Massive reorganizations of the ocean-atmosphere system, the authors argue, are the key events that link cyclic changes in the earth's orbit to the advance and retreat of ice sheets

by Wallace S. Broecker and George H. Denton

Eight times within the past million years, something in the earth's climatic equation has changed, allowing snow in the mountains and the northern latitudes to remain where it had previously melted away. The snow compacted into ice, and the ice built up into glaciers and ice sheets. Over tens of thousands of years, the ice sheets reached thicknesses of several kilometers; they planed, scoured and scarred the landscape as far south as central Europe and the midwestern U.S. And then each glacial cycle came to an abrupt end. Within a few thousand years, the ice sheets shrank back to their present-day configurations.

Over the past 30 years, evidence has mounted that these glacial cycles are ultimately driven by astronomical factors: slow, cyclic changes in the eccentricity of the earth's orbit and in the tilt and orientation of its spin axis. By altering the intensity of the seasons, the astronomical cycles somehow tip the balance between glacial buildup and glacial retreat. But what is the link between astronomy and the ice ages? How are the seasonality changes leveraged into global changes in climate?

Any answer must contend with the vast array of evidence that has accumulated about the nature, timing and extent of the climatic shifts that accompanied ice buildup and retreat. Many workers have proposed that the

seasonality changes act directly on the ice sheets of the Northern Hemisphere. A reduction in summer sunshine allows ice to build up, and an increase melts it away; the ice in turn alters the earth's climate. In contrast, we think the ice sheets were a consequence of broader climatic events. By altering patterns of evaporation and rainfall, the changes in seasonal intensity appear to have caused the ocean and atmosphere (a single, coupled system) to flip from one mode of operation to another, very different mode. With each flip, ocean circulation changed and heat was carried around the globe differently, the properties of the atmosphere were altered, climate changed—and the ice sheets grew or shrank.

Our proposal is not a rejection of the astronomical theory of the ice ages but an extension of it. The hypothesis was first proposed in 1842, just a few years after the Swiss-American naturalist Louis Agassiz argued that polished and scarred rocks and heaps of detritus in the Alps recorded some past age of glaciers. In that year the French mathematician Joseph A. Adhémar suggested that astronomically driven changes in the intensity of the seasons might periodically trigger glaciation.

The Yugoslav astronomer Milutin Milankovitch refined and formalized the hypothesis in the 1920's and 1930's. The astronomical pacemaker he advocated has three components, two that change the intensity of the seasons and a third that affects the interaction between the two driving factors. The first is the tilt of the earth's spin axis. Currently about 23.5 degrees from the vertical, it fluctuates from 21.5 degrees to 24.5 degrees and back every 41,000 years. The greater the tilt is, the more intense seasons in both hemispheres become: summers get hotter and winters colder.

The second, weaker factor control-

ling seasonality is the shape of the earth's orbit. Over a period of 100,000 years, the orbit stretches into a more eccentric ellipse and then grows more nearly circular again. As the orbital eccentricity increases, the difference in the earth's distance from the sun at the orbit's nearest and farthest points grows, intensifying the seasons in one hemisphere and moderating them in the other. (At present the earth reaches its farthest point during the Southern Hemisphere winter; as a result, southern winters are a little colder—and summers a little warmer—than their northern counterparts.)

A third astronomical fluctuation governs the interplay between the tilt and eccentricity effects. It is the precession, or wobble, of the earth's spin axis, which traces out a complete circle on the background of stars about every 23,000 years. The precession determines whether summer in a given hemisphere falls at a near or a far point in the orbit—in other words, whether tilt seasonality is enhanced or weakened by distance seasonality. When these two controllers of seasonality reinforce each other in one hemisphere, they oppose each other in the opposite hemisphere.

WALLACE S. BROECKER and GEORGE H. DENTON bring diverse interests to their study of ice ages. Broecker got his Ph.D. at Columbia University in 1958 and has pursued his career there. He is now professor of geochemistry at the Lamont-Doherty Geological Observatory of Columbia University. In addition to ancient climates, he follows research interests in ocean chemistry, isotope dating and environmental science. Denton is professor of geology at the University of Maine. After earning a Ph.D. at Yale University in 1965, he did postdoctoral work at the University of Stockholm and then moved to Maine. He has spent 36 seasons in the field studying the timing and extent of glacial advances, 22 of them in Antarctica and elsewhere in the Southern Hemisphere.

ICE FIELD IN PATAGONIA ends in a deep glacial lake. Such Southern Hemisphere glaciers have grown and shrunk in concert with the great northern ice sheets, according to radiocarbon dating of vegetation (such as the trees in the foreground) that was overwhelmed by advancing glaciers or that took root after their retreat. The timing is a puzzle because the intensity of summer sunshine, which is thought to influence ice growth, changes on quite different schedules at middle latitudes in the two hemispheres.

Milankovitch calculated that these three factors work together to vary the amount of sunshine reaching the high northern latitudes in summer over a range of some 20 percent—enough, he argued, to allow the great ice sheets that advanced across the northern continents to grow during intervals of cool summers and mild winters. For

many years, however, the lack of an independent record of ice-age timing made the hypothesis untestable.

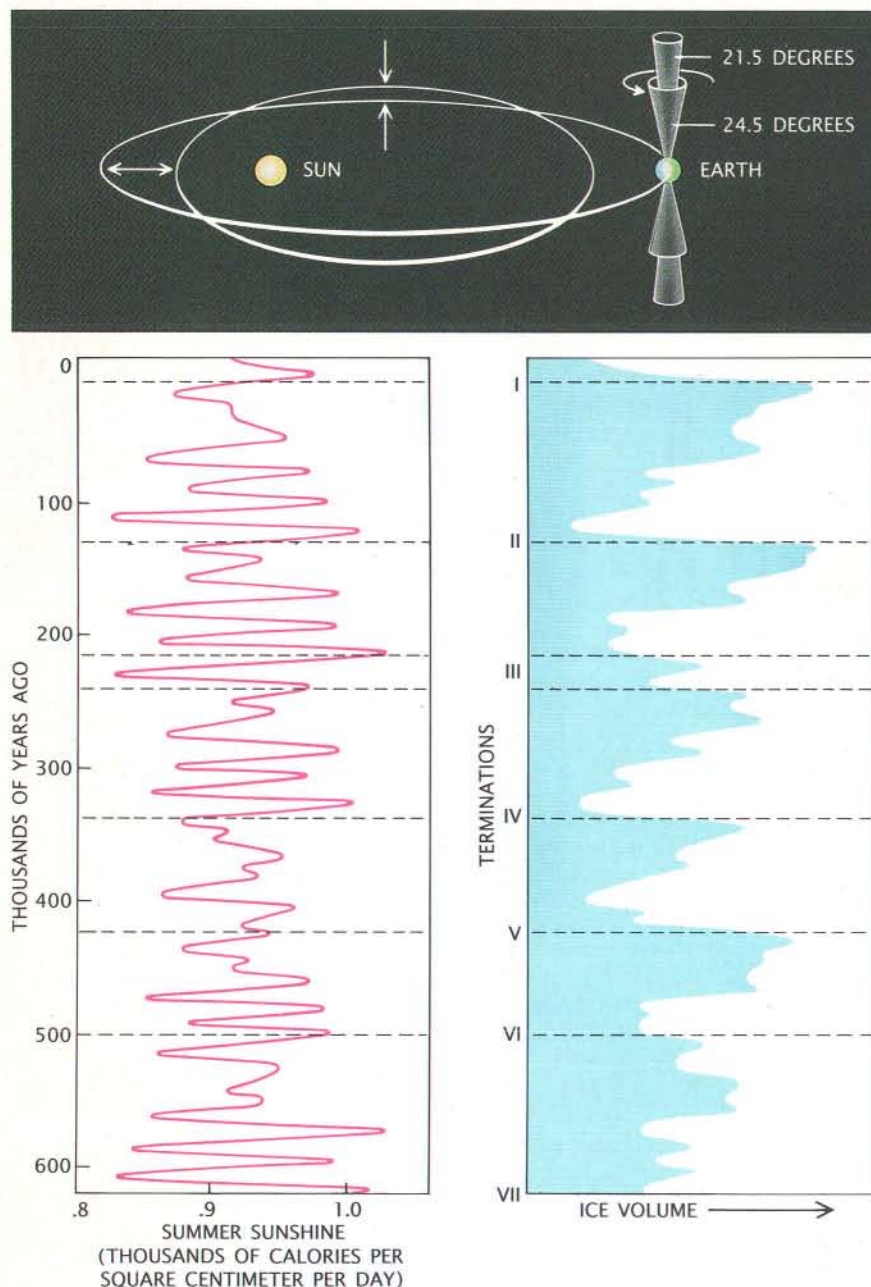
In the early 1950's Cesare Emiliani, working in Harold C. Urey's laboratory at the University of Chicago, produced the first complete record of the waxings and wanings of past glaci-

ations. It came from a seemingly odd place, the sea floor. Single-cell marine organisms called foraminifera house themselves in shells made of calcium carbonate. When the foraminifera die, sink to the bottom and contribute to the sea-floor sediments, the carbonate of their shells preserves certain characteristics of the seawater they inhabited. In particular, the ratio of a heavy isotope of oxygen (oxygen 18) to ordinary oxygen (oxygen 16) in the carbonate preserves the ratio of the two oxygens in the water molecules.

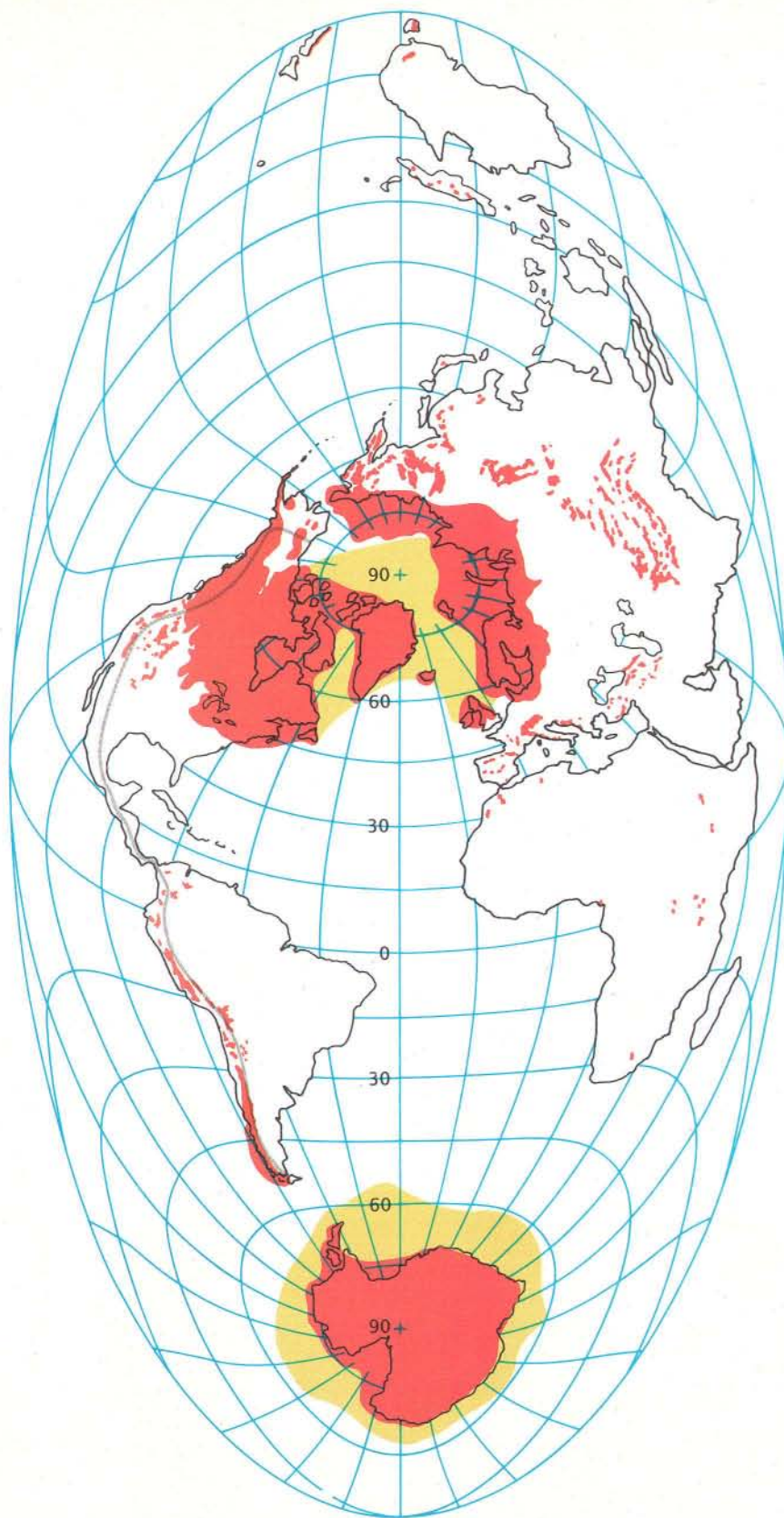
It is now understood that the ratio of oxygen isotopes in seawater closely tracks the proportion of the world's water that is locked up in glaciers and ice sheets. A kind of meteorological distillation accounts for the link. Water molecules containing the heavier isotope tend to condense and fall as precipitation a tiny bit more readily than molecules containing the lighter isotope. Hence, as water vapor evaporated from warm oceans moves away from the source, its oxygen 18 preferentially returns to the oceans in precipitation. What ultimately falls as snow on ice sheets and mountain glaciers is relatively depleted of oxygen 18. As the oxygen 18-poor ice builds up, the oceans become relatively enriched in the isotope. The larger the ice sheets grow, the higher the proportion of oxygen 18 becomes in seawater—and hence in the sediments.

Analyzing cores drilled from sea-floor sediments, Emiliani found that the isotopic ratio rose and fell in rough accord with the cycles Milankovitch had predicted. Since that pioneering observation, oxygen-isotope measurements have been made on hundreds of cores. A chronology for the combined record enabled James D. Hays of Columbia University, John Imbrie of Brown University and Nicholas Shackleton of the University of Cambridge to show in 1976 that the record contains the very same periodicities as the orbital processes.

Over the past 800,000 years, the global ice volume has peaked every 100,000 years, matching the period of the eccentricity variation. In addition, "wrinkles" superposed on each cycle—small decreases or surges in ice volume—have come at intervals of roughly 23,000 and 41,000 years, in keeping with the precession and tilt frequencies. Imbrie, working with a group called SPECMAP, later strengthened the case for the astronomical theory even more when he showed that the amplitude of the shorter-period signals has varied exactly as one would expect if the signals were be-

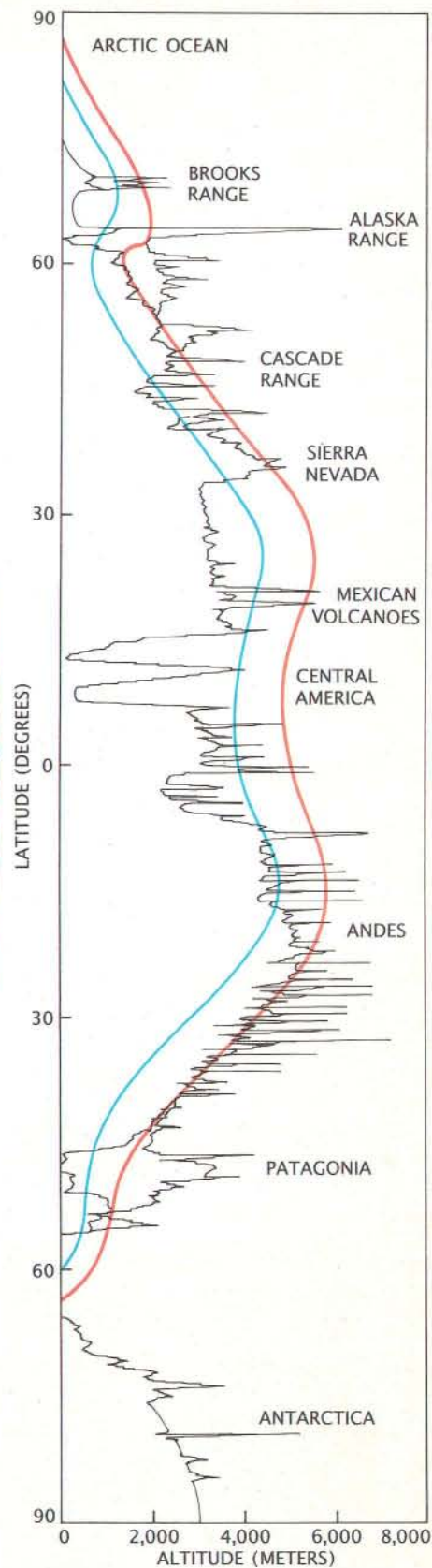


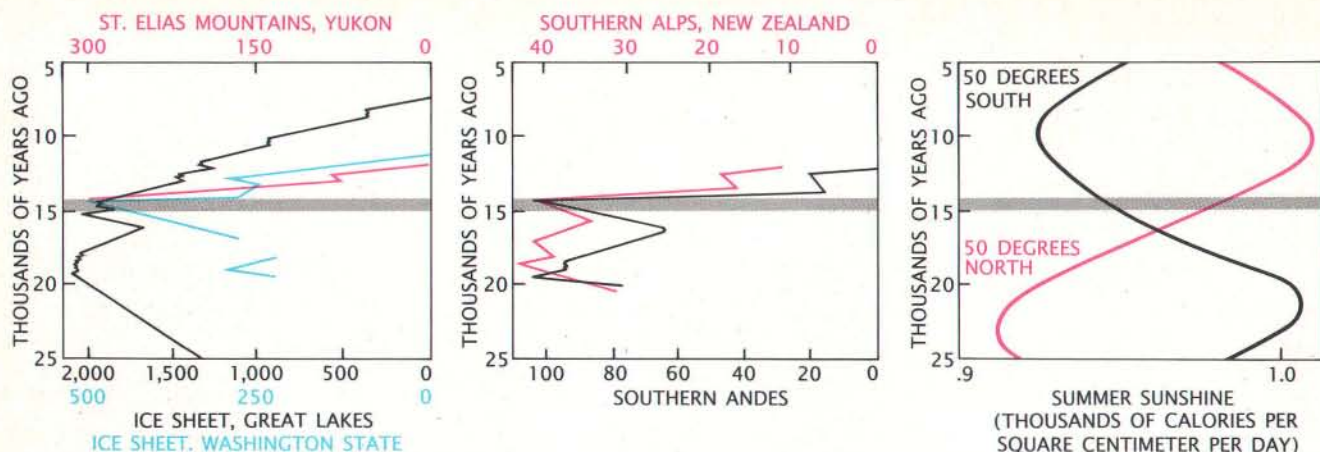
ASTRONOMICAL CYCLES (top) are the pacemaker of glaciation. The cycles—23,000 to 100,000 years in length—affect the eccentricity of the earth's orbit, the orientation of its spin axis (which slowly traces out a cone in space) and the tilt of the axis (which affects the width of the cone). The effect of the changes on the intensity of summer sunshine at high northern latitudes is shown at the left. The curve at the right indicates the volume of the earth's ice sheets, determined from isotopic studies of sea-floor sediments. Ice volume climbs gradually for about 100,000 years and then falls abruptly in ice-age terminations that correspond to episodes of increasing summer sunshine at northern latitudes. (Seasonality varies differently in the Southern Hemisphere, which suggests that northern seasonality must be what drives ice ages.)



ICE SHEETS AND MOUNTAIN GLACIERS expanded in both hemispheres during the last ice age. The map (an unusual equal-area projection) shows the extent of land ice (red) and sea ice (yellow) on all the continents at peak glaciation some 19,500 years ago. (Land ice extended beyond some present coastlines

because the sea level was lower.) The graph traces the average elevation of mountain snow lines on the American cordillera, plotted along the north-south transect indicated on the map. Ice-age snow lines (blue line) were about 1,000 meters lower than snow lines are today (red), regardless of latitude.





TIMING of glacial retreat was identical in the Northern Hemisphere (left) and in the Southern Hemisphere (center). The graphs give the extent of mountain glaciers and ice sheets from their source regions (in kilometers) and show that in ev-

ery case dramatic retreat began 14,000 years ago. Changes in seasonal intensity could not have driven the retreat directly, because even though northern summers were getting stronger, summers in the Southern Hemisphere were weakening (right).

ing modulated by distance seasonality.

To be sure, there were loose ends. The 100,000-year variation has a much weaker effect on seasonal sunshine than the shorter cycles do, and yet it apparently sets the fundamental frequency of glaciation. The shorter cycles emerge only in the wrinkles in the isotopic record. What is more, the calculated seasonality cycles rise and fall smoothly, but the ice curve is saw-toothed: the ice grows episodically for nearly 100,000 years and then crashes in a few thousand, in a period of strengthening northern summers.

Workers have sought answers to both puzzles in the physics of the ice sheets and the underlying rock, which sinks under the weight of the ice. For example, William R. Peltier and William T. Hyde of the University of Toronto have built a theoretical model that incorporates assumptions about how the bedrock sinks and that closely reproduces both the dominance of the 100,000-year cycle and the rapid retreat of the ice. In the model, it takes nearly 100,000 years for an ice sheet to reach a critical size, at which point the ductile rock below the earth's crust begins to flow rapidly and allows the burdened crust to sink. The surface of the ice sheet drops; warmed by the lower elevation, the ice can melt rapidly when the shorter-period cycles bring the next episode of strong northern summers.

Peltier and Hyde's model, like many other models, assumes that Northern Hemisphere seasonality changes drive glacial advance and retreat directly, with bedrock response shaping each cycle and setting its length. Yet the assumption suffers a crucial problem: glaciers grew and

retreated in the Southern Hemisphere as well. Studies by geologists, including the late John H. Mercer of Ohio State University and Stephen C. Porter of the University of Washington, show that during the last ice age, climate changed at the same times and by comparable amounts in the middle latitudes of the Southern Hemisphere—even though seasonality there varies on a quite different schedule.

They and others have found, for example, that during the last ice age the earth's mountain glaciers also expanded. The evidence—from the heaps of debris plowed up by the glaciers, known as moraines—is as clear in the tropics (New Guinea, Hawaii, Colombia and East Africa) and the southern temperate latitudes (Chile, Tasmania and New Zealand) as it is in northern temperate latitudes (the Cascades, the Alps and the Himalayas). On all the mountains studied so far, regardless of geographic setting or precipitation rate, the snow line descended by about one kilometer, corresponding to a drop in temperature of about five degrees Celsius.

Where organic material was trapped in the moraines, radiocarbon dating shows that the glaciers advanced and retreated on the same schedule. They fluctuated near their maximum extent between about 19,500 and 14,000 years ago, about the same time as the glaciation of northern continents peaked. Then, just as the northern ice sheets began to shrink, the mountain glaciers underwent a dramatic retreat that sharply reduced their size by about 12,500 years ago.

How could changes in summer sunshine at the latitude of Iceland have caused glaciers to grow and retreat in New Zealand and the southern An-

des? If orbital cycles do indeed drive glacial cycles by acting directly on northern ice sheets, the response to seasonality changes in the high northern latitudes must be strong enough to override the effects of the very different changes in the Southern Hemisphere. One possibility is that the northern ice sheets themselves translate Northern Hemisphere seasonality into climatic change around the world.

Two links between the northern ice sheets and ice growth worldwide have been proposed, but neither one bears up well under scrutiny. One invokes sea level, which would have dropped as the growth of the northern ice locked up much of the world's water. Since glaciers can grow only on land, the drop in sea level might have allowed southern glaciers to expand onto the exposed continental shelves even without a global change in temperature. Later, when the northern ice sheets melted, the rise in sea level might have broken up the margins of the Southern Hemisphere glaciers, forcing them to retreat. The explanation is plausible only for Antarctica, however, because most mountain glaciers do not approach the sea.

The second proposal relies on the high albedo, or reflectivity, of the vast northern ice sheets. By reducing the absorption of sunlight by the planet as a whole, the ice might have led to global cooling and allowed glaciers to grow at southern latitudes. Yet computer climate models show that the albedo effects of Northern Hemisphere ice sheets should be confined to northern latitudes. Also, if ice albedo does drive global climatic change, one would expect to find a pronounced north-to-south gradient in the mountain-glacier record,

with mountains adjacent to the northern ice sheets recording the greatest snow-line lowering and the Andes, say, showing very little change. No such gradient is seen.

Any causal link between the ice sheets and global climatic change also must contend with the timing of the mountain-glacier retreat. Both the northern ice sheets and the mountain glaciers began their retreat from the last glacial maximum at the same time, about 14,000 years ago. The continental glaciers took about 7,000 years to melt away, whereas the mountain glaciers shrank much more quickly. The disparity suggests that the northern ice sheets cannot be calling the tune for climate over the rest of the earth.

If the ice sheets themselves cannot link the astronomical cycles to the climatic shifts, what can? Clues come from core samples drilled from depths of as much as two kilometers in the ice that still blankets Greenland and Antarctica. The first thing the ice cores offer is confirmation of the global and synchronous character of ice-age climatic changes.

The oxygen 18 content of glacial ice is depleted in general, but the exact content records the local temperature at the time the ice was laid down. (The colder a parcel of air becomes, the more of its water vapor is likely to have fallen already in precipitation, reducing the oxygen 18 content of the remaining vapor.) Isotopic studies of the Greenland and Antarctic cores show that during the last glaciation both poles cooled—to as much as 10 degrees C below today's temperatures—and warmed in step.

The ice also revealed something much more intriguing. Groups led by Hans Oeschger of the University of Bern and Claude Lorius of the Laboratory of Glaciology and Geophysics of the Environment, near Grenoble, measured the carbon dioxide content of the tiny bubbles of ancient air trapped in the ice. They found that during the last glaciation the carbon dioxide concentration of the atmosphere was about two thirds of its interglacial level. The carbon dioxide curve pointed to a missing ingredient in the climatic recipe: the ocean.

Only a major shift in the ocean's operation could account for such a dramatic change in atmospheric composition. After all, the oceans hold 60 times as much carbon dioxide as the atmosphere; because the gas readily diffuses between the ocean surface and the atmosphere, its concentration

in surface waters regulates the atmospheric concentration.

Living things in turn control the surface-water concentration, by acting as a biological pump that transfers carbon dioxide from the surface to the ocean depths. In the course of photosynthesis, the tiny green plants of the ocean's sunlit upper layers capture dissolved carbon dioxide to form organic tissue. Some of the plant matter, as well as animal tissue nourished by the plants, eventually sinks into the deep sea, where bacteria oxidize it back to carbon dioxide. Thus, the gas is continuously pumped into the abyss, together with nutrients such as phosphate and nitrate.

The efficiency of this pump depends not only on the surface community's population and species but also on vertical mixing patterns. The exact link between pumping efficiency and ocean circulation is controversial, but one can imagine, for example, that if the mixing of deep waters with the surface is slowed, surface plant life will have more time to deplete the shallow water of carbon dioxide before more of the gas is stirred up from the depths. During glacial time, some combination of altered mixing and changes in ecology must have made the biological pump more efficient.

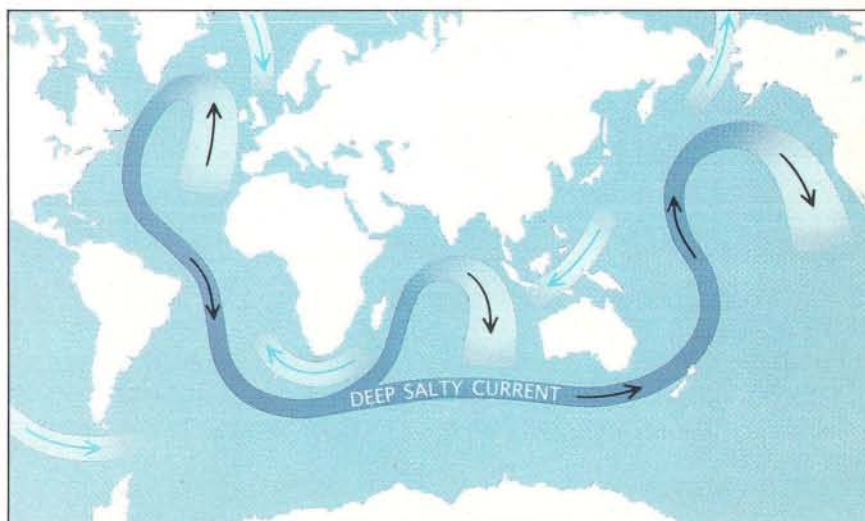
The first indications that the ice-age ocean did operate differently came from fossil evidence: changes in the populations of micro-

organisms that inhabit water masses of specific temperature and salinity, studied by William F. Ruddiman and Andrew McIntyre of Columbia University and by Detmar F. Schnitker of the University of Maine. More recently a geochemical technique pioneered by Edward A. Boyle of the Massachusetts Institute of Technology provided dramatic and direct confirmation that the ocean circulated differently during the last glaciation.

Boyle discovered that, for unknown reasons, the distribution of cadmium in today's oceans closely matches that of phosphate and nitrate nutrients. Because the cadmium ion has the same charge and size as calcium, Boyle guessed that cadmium might substitute for calcium in the calcium carbonate of foraminiferal shells. If it does, measurements of cadmium in shells from sediment cores might reveal the distribution of nitrate and phosphate in the glacial ocean.

Boyle's intuition proved correct when he found that foraminifera in the present-day ocean do incorporate cadmium in a constant proportion to its abundance in seawater. He then measured cadmium in sediment cores. The result was exciting: a key signature of the Atlantic's present-day circulation was missing during glacial time, until about 14,000 years ago.

Currently the Atlantic's deep water contains only about half as much phosphate and nitrate as the deep waters of the Pacific and Indian oceans. The



DEEP SALTY CURRENT threads the world's oceans, compensating for the transport of water vapor by the atmosphere. (Light blue arrows indicate shallow return flow.) The current originates in the North Atlantic, where northward-flowing warm water that is unusually saline (and therefore dense) because of excess evaporation is chilled, which increases its density further. It sinks into the abyss and flows southward, out of the Atlantic. Most of the salty water that is supplied by this Atlantic "conveyor" mixes upward in the Pacific, making up for excess precipitation there. The Atlantic conveyor—and probably the entire system—was disrupted during glacial time.

low nutrient content reflects the water's recent sojourn near the surface (where biological activity depletes the nutrients). Every winter at about the latitude of Iceland, water of relatively high salinity, flowing northward at intermediate depths (perhaps 800 meters), rises as winds sweep the surface waters aside. Exposed to the chill air, the water releases heat, cooling from 10 degrees C to two degrees. The water's high salinity together with the drop in temperature makes it unusually dense, and it sinks again, this time all the way to the ocean bottom.

The formation of the North Atlantic deep water, as it is called, gives off a staggering amount of heat. Equal to about 30 percent of the yearly direct input of solar energy to the surface of the northern Atlantic, this bonus of heat accounts for the surprisingly mild winters of Western Europe. (The warming is often mistakenly ascribed to the Gulf Stream, which ends well to the south.) The magnitude of the vertical circulation is also immense, averaging 20 times the combined flow of all the world's rivers. Indeed, much of

the deep water in the world's oceans ultimately originates here. From its source the water floods the deep Atlantic, curves around the southern tip of Africa and joins the deep current that circles Antarctica and distributes deep water to the other oceans.

As the deep water ages and travels away from the site of its formation, it collects sinking phosphate and nitrate, which results in a gradient of increasing nutrient levels. By measuring the cadmium content of foraminifera that lived near the bottom, Boyle found that during glacial time the nutrients were more uniformly distributed through the depths of the world's oceans. In addition, the concentration in the glacial Atlantic peaked in the deepest parts rather than at intermediate depths, as it does today.

These results bore out the implication of the earlier microfossil studies. The Atlantic "conveyor," which releases vast quantities of heat to the North Atlantic and sends immense volumes of water into the abyss, was shut down until the last ice age ended 14,000 years ago. In the absence of this key component, worldwide ocean circula-

tion must have looked very different.

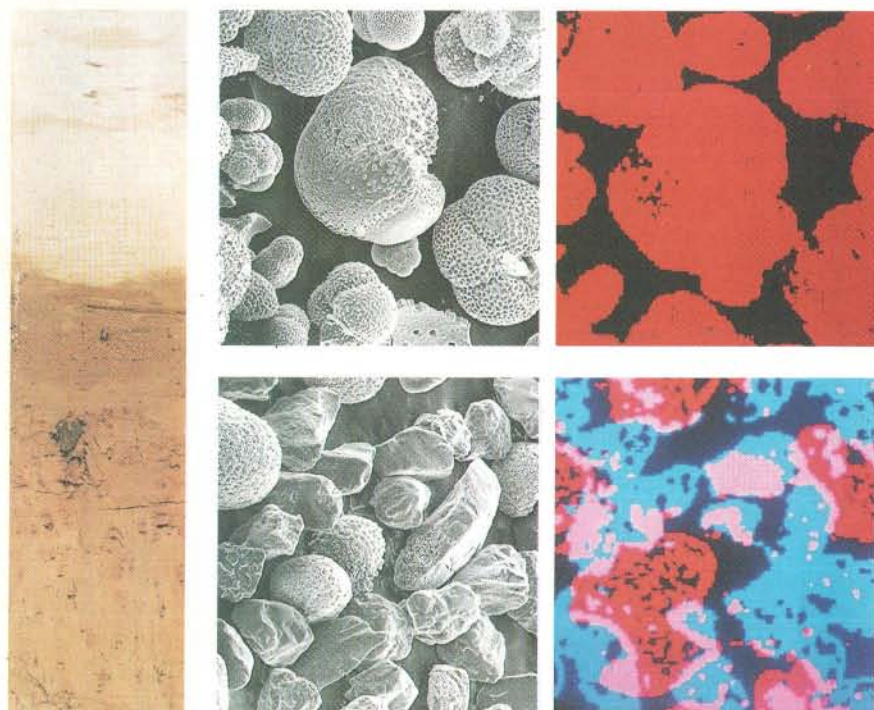
The sea and land evidence together points to a simultaneous change in the operation of the ocean and the atmosphere 14,000 years ago. The pattern of ocean circulation shifted dramatically; glaciers in both hemispheres began retreating, signaling global warming; and the carbon dioxide content of the atmosphere started to rise to interglacial levels. We think these events indicate a major reorganization of the joint ocean-atmosphere system—a jump from a glacial mode of operation to an interglacial mode. Indeed, we believe that abrupt jumps among several ocean-atmosphere modes may underlie glacial cycles in general.

We propose that changes in seasonality are the ultimate causes of these mode shifts. Although we can suggest no simple mechanisms linking seasonality, the ocean-atmosphere system and global climate, we can offer some insights.

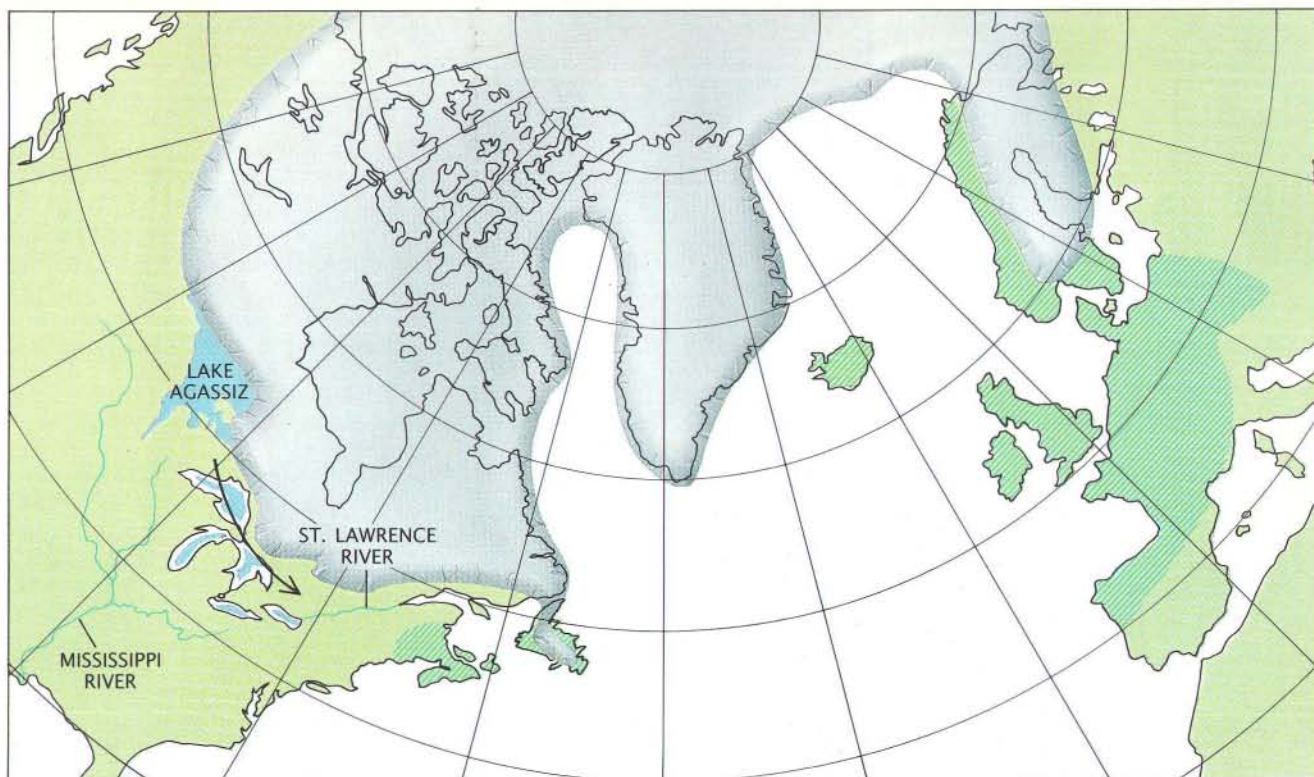
The atmosphere, which would certainly feel the effects of seasonality changes, strongly influences the circulation of the ocean. The link involves the distribution of salt. Prevailing winds transfer water evaporated from one part of the ocean to another region, where it falls as precipitation. The transport of vapor leaves a heritage of salt in the first region and dilutes the salinity of the second.

Now, the tendency of surface waters to sink into the depths and initiate a vertical conveyor belt like that of the North Atlantic depends on their density. Density reflects both temperature and salinity, but salinity is the decisive factor. (Surface water cools almost to the freezing point throughout the high latitudes in winter, but only where it is unusually saline does it sink into the abyss.) The system has a built-in nonlinearity: a gradual shift in atmospheric circulation, by changing salinity in regions such as the North Atlantic, could dramatically alter the global circulation pattern. Indeed, the Atlantic conveyor appears to be the most vulnerable part of the system, which may explain why it is Northern Hemisphere seasonality that drives global climatic changes.

A climatic event called the Younger Dryas, which took place several thousand years after the glaciers started to retreat, provides a smoking gun for this part of our case. It vividly illustrates the link between the transport of fresh water—in this case liquid water and not vapor—and ocean circulation. About 11,000 years ago the re-



SEDIMENT CORE (left) from the North Atlantic testifies to an abrupt change in circulation at the end of the penultimate glaciation about 128,000 years ago. The transition (identified by Gerard C. Bond of Columbia University) spans a few millimeters and represents about 50 years. A scanning electron micrograph of coarse material from the dark sediments (bottom) reveals abundant rock fragments, rich in silicon (blue in an X-ray map), presumably dropped by melting icebergs. The light-colored sediments (top) include almost no rock and are made up mainly of shells, rich in calcium (red), from marine organisms that inhabit warm waters. (Shells in the dark sediments came from cold-water species.) The sudden revival of the Atlantic conveyor must have warmed the surface, eliminating icebergs and altering ecology.



DIVERSION OF MELTWATER during the retreat of the North American ice sheet some 11,000 years ago may explain the 1,000-year cold spell known as the Younger Dryas. Lake Agassiz, fed by meltwater, had been draining down the Mississippi River to the Gulf of Mexico. When the retreat of the ice opened a channel to the east, however, the water flooded

across the region of the Great Lakes to the St. Lawrence River (arrow). The influx of fresh water to the North Atlantic diluted the salinity of surface water, reducing its density and preventing it from sinking. The Atlantic conveyor was shut down: warm water could no longer flow northward, and a broad region around the North Atlantic was chilled (hatched area).

treat of the glaciers was well under way, and temperatures had risen to their interglacial levels. Suddenly, in as little as 100 years, northern Europe and northeastern North America reverted to glacial conditions. Pollen records show that the forests that had colonized postglacial Europe gave way to arctic grasses and shrubs (including the Dryas flower, for which the period is named), and the Greenland ice core records a local cooling of six degrees C. About 1,000 years later, this cold spell ended abruptly—in as little as 20 years, recent work by Willi Dansgaard of the University of Copenhagen suggests.

Boyle's cadmium measurements, together with the record of surface-water foraminifera in the North Atlantic, tell what happened. Both indicators return to their glacial state at the onset of the Younger Dryas. The conveyor belt had shut down once again. Deep-water formation had stopped, and so the warm intermediate-depth water that supplies Europe's bonus of heat could no longer flow northward. The chill over the region was dispelled only when the conveyor began running again 1,000 years later.

A massive influx of fresh water from the melting North American ice sheet seems to have killed the conveyor and precipitated the Younger Dryas. The ice sheet started shrinking 14,000 years ago; for the 7,000 years it took to melt away, it must have released fresh water at about the same rate as today's Amazon River. At first nearly all the meltwater from the southern edge of the massive ice sheet flowed down the Mississippi River to the Gulf of Mexico. About 11,000 years ago, however, a major diversion sent meltwater in torrents down the St. Lawrence River to the Atlantic.

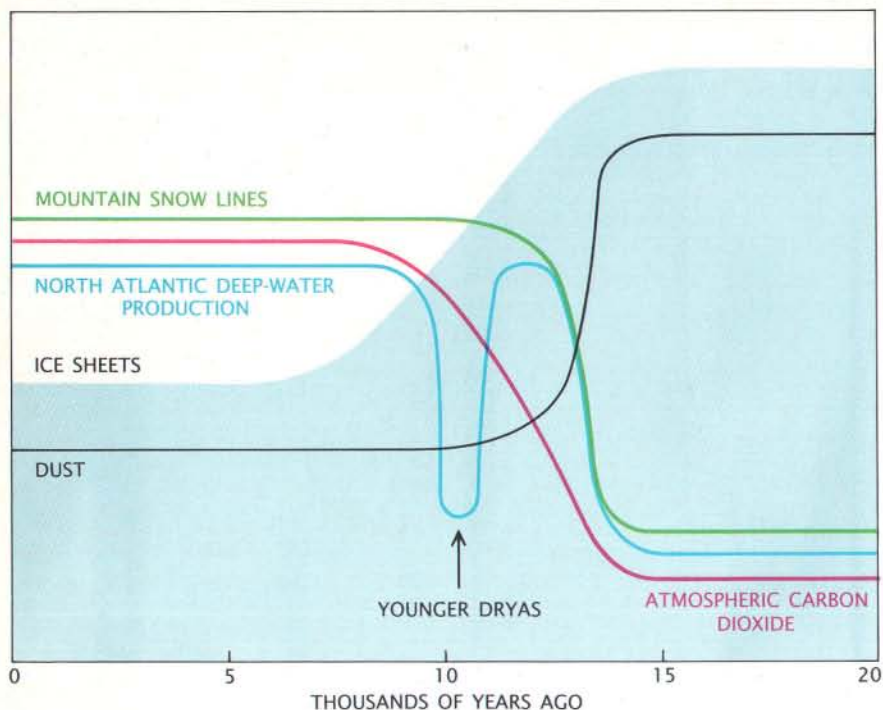
A vast clearinghouse for meltwater, known as Lake Agassiz, had formed in the bedrock depression at the edge of the retreating ice sheet in what is now southern Manitoba. Until 11,000 years ago the lake, larger than any of the existing Great Lakes, had overflowed a bedrock lip to the south and drained down the Mississippi. Then the retreat of the ice opened a channel to the east. The water level in Lake Agassiz dropped by 40 meters as water flowed across the region of the Great Lakes and down the St. Lawrence.

Foraminifera from surface waters of

the Gulf of Mexico record this diversion. Their oxygen 18 content had been anomalously low, reflecting the oxygen 16-rich meltwater discharging from the Mississippi. About 11,000 years ago the isotopic ratio increased abruptly as the Lake Agassiz diversion shut off the meltwater flow to the Gulf.

The meltwater, meanwhile, poured into the North Atlantic close to the site of deep-water formation. There it reduced the salinity of surface waters (and hence their density) by so much that, in spite of severe winter cooling, they could not sink into the abyss. The conveyor belt stayed off until 1,000 years later, when a lobe of ice advanced across the western end of the Lake Superior basin and once again blocked the exit to the east. Lake Agassiz rose again by 40 meters, diverting the meltwater back down the Mississippi. The conveyor belt was reactivated, and Europe warmed up again.

The Younger Dryas links freshwater flow, ocean circulation and climate—but only regional climate. Only around the North Atlantic did the episode bring a sharp cooling; elsewhere its effects were slight



END OF THE LAST ICE AGE brought global changes, summarized here, that began at the same time about 14,000 years ago even though they proceeded at different rates. The circulation of the North Atlantic shifted abruptly from glacial to interglacial conditions (with a brief relapse during the Younger Dryas cold snap) as deep-water production resumed. At the same time, the amount of dust in the atmosphere dropped and the concentration of carbon dioxide started to increase. The shifts may have been part of a larger reorganization of the ocean and atmosphere that warmed the planet and caused mountain glaciers and ice sheets to start retreating.

or absent. Unlike the glaciations, the Younger Dryas affected only the transport of heat (from low latitudes to the North Atlantic) and not the global climate. How could a change in ocean-atmosphere operation during the ice ages have cooled the world as a whole?

The Greenland and Antarctic ice cores suggest part of an answer. The lower level of atmospheric carbon dioxide they record for the last glaciation would certainly have contributed to the cooling: carbon dioxide is a greenhouse gas that warms the earth's surface by trapping solar energy. Computer climate simulations suggest, however, that the global cooling caused by the observed drop in carbon dioxide would be at most two degrees C—less than half of what is recorded in the mountain glaciers.

Two other changes recorded in the ice cores must also have contributed. Ice-age air contains only half the post-glacial level of methane. Methane, too, is a greenhouse gas, although the ice-age cooling attributable to reduced methane amounts to just a few tenths of a degree. In addition, dust is about 30 times as abundant in glacial-age ice as in more recent layers, confirming evidence from other sites that the

ice-age atmosphere was exceedingly dusty. Dust, too, could have contributed to the cooling, by reflecting sunlight. Unfortunately, its effect is hard to quantify.

The dustiness and low methane content of the ice-age air do suggest that the glacial mode of ocean-atmosphere operation had imposed a dry climate. Dust, after all, blows from areas where vegetation is sparse, whereas methane is produced in swamps. Dry conditions (which are also recorded in ice-age landforms, such as sand dunes, and in pollen deposits) would have had their own effect on global temperatures. Temperature falls more rapidly with increasing altitude in a drier atmosphere; hence, the drying could have contributed to the depression of mountain snow lines.

Even added together, the impacts of carbon dioxide, methane, dust and drying may come up short in accounting for the temperature difference between the glacial and the interglacial planet. What else could have contributed? One possibility is that the ocean-atmosphere reorganization changed the characteristics of clouds and made them more reflective.

Clearly, our account of how changes in ocean-atmosphere operation could have cooled the planet is incomplete. Moreover, since we appeal to Northern Hemisphere seasonality to pace these mode shifts, we encounter the same problem faced by other theorists: Why is the 100,000-year astronomical cycle dominant when it is the weakest of the three? Perhaps ice-sheet growth has a feedback effect on atmospheric circulation. The ocean-atmosphere system might become most susceptible to a mode shift once the ice sheets reached a critical size—which might take 100,000 years.

Still, much recent evidence favors our basic proposal: transitions between glacial and interglacial conditions represent jumps between two stable but very different modes of ocean-atmosphere operation. If the earth's climate system does jump between quantized states, like the electrons around an atom, all climate indicators should register a transition simultaneously. In this regard, the evidence from the end of the last ice age is most impressive. The warming of North Atlantic surface waters, the onset of melting in the northern ice sheets and the mountain glaciers of the Andes, the reappearance of trees in Europe and changes in plankton ecology near Antarctica and in the South China Sea—all took place between 14,000 and 13,000 years ago.

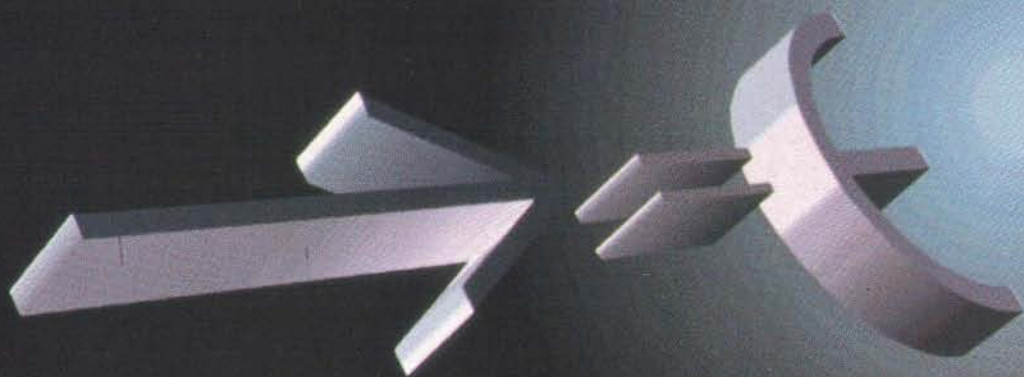
If the global climate system does prove to have quantized states, climatologists will have gained new insight into the way astronomical forcing, acting mainly in high northern latitudes, could transform climate worldwide. They will also have new cause for concern over the earth's climatic future. Just as 14,000 years ago the earth was feeling the gradual forcing effect of stronger northern summers, so now it is subject to gradual forcing as human activity releases carbon dioxide and other greenhouse gases into the atmosphere. Will the climate system again respond abruptly, by flipping to an entirely new mode?

FURTHER READING

THE OCEAN. Wallace S. Broecker in *Scientific American*, Vol. 249, No. 3, pages 100-112; September, 1983.

ICE AGES: SOLVING THE MYSTERY. John Imbrie and Katherine O. Imbrie. Harvard University Press, 1986.

THE ROLE OF OCEAN-ATMOSPHERE REORGANIZATIONS IN GLACIAL CYCLES. Wallace S. Broecker and George H. Denton in *Geochimica et Cosmochimica Acta*, Vol. 53, No. 10, pages 2465-2501; October, 1989.



Our force is your energy

With every product and every activity, Olivetti has just one aim: to concentrate the benefits and the full potential of the power of technology within the hands of the user.

This means making information science more useful and more useable, in more ways, for more people than any other company involved in information technology.

It is from you, the user, that we get all the best ideas for improving ourselves and everything we do is for you. "Our force is your energy".

olivetti

SCIENCE AND BUSINESS

Computer Babble

Manufacturers move slowly to adopt open systems

About five years ago managers at Eastman Kodak made a vexing observation: the multitude of computers at the disposal of their engineers did not always help. Many investigators had to spend half of their time writing software tools simply to communicate with other computers at Kodak or even to run their research problems on the machines in the first place.

Kodak's difficulties were typical of those faced by the corporations that had snapped up fast, powerful systems such as workstations. To achieve their high-performance goals, these machines relied on specialized software codes. But the software differences meant that various species of computers were often incompatible.

Now "we are just turning the corner," says Charles J. Gardner, a Kodak manager. Slowly and reluctantly, computer manufacturers have begun developing software standards for every layer of code—from the operating system that sends commands to the machine hardware to the interface a user sees on screen. A system complying with these standards would be "open." Open systems could exchange data and files easily; software could be easily ported (or moved) to any open system, from a personal computer to a supercomputer. Yet the work is far from complete. A host of technical hurdles remain. Moreover, it is clear that computer makers will pay a price for the changes: the transition to open systems may hurt profit margins and increase competition.

In the past, vendors of high-powered computers flaunted their differences. Customers who installed a system either implicitly committed themselves to buying future versions of it or accepted the necessity of one day having to spend big money to move projects from the installed system to a competitor's new machines.

Finally, customers began to object. Beginning in the early 1980's, they banded together, forming an alphabet soup of standards groups to work toward establishing uniform protocols. Along with the traditional committees sponsored by the Institute of Electrical and Electronics Engineers (IEEE) as



Open computers, designer drugs, entrepreneur-san, the Federal Reserve

well as the American National Standards Institute, there were ISO, OSI, SPAG, X/Open, MAP, TOP, COS and eventually OSF and UNIX International.

Meanwhile, on its own, Sun Microsystems, a maker of high-speed workstations in Mountain View, Calif., was proving that manufacturers could succeed by designing new machine architectures able to run existing software. Sun also licensed its technology widely. As a result, Sun's chief competitor, Apollo, soon found itself supporting not only its own technology but also the de facto industry standards developed by Sun.

When the U.S. government began requiring that the computer systems it bought conform to the emerging industry standards, most manufacturers bowed to the inevitable and joined one or more standards organizations. (Apple remains the most aloof.) These groups then tackled various layers of software codes.

Much progress has been made in developing standards that enable machines to swap data and files easily. Even so, "getting enough vendors to build to the same set of standards" has been tough, Gardner says.

In contrast, the search for a common operating system—only one of the many layers of software between the machine hardware and such software applications as spreadsheet programs—has been marked by discord and remains unresolved. A few years ago some variation of UNIX, an operating system developed at AT&T in the 1970's, seemed likely to prevail—but which one? When AT&T and Sun announced in 1987 that they would together update AT&T's System V variant, other companies worried that Sun would gain an unfair advantage.

A battle ensued. Seven vendors, including IBM, Digital Equipment and

Hewlett-Packard, formed the Open Software Foundation (OSF) to develop "an open operating-system software environment." AT&T, Sun and other vendors countered with UNIX International, created to direct "the evolution of UNIX System V, the industry-standard open operating system." This past November UNIX International introduced System V Release 4 operating system, which merged AT&T's UNIX with two other dominant strains; OSF hopes to ship OSF/1 in late 1990.

The two organizations have "created enormous confusion in the industry," charges Eric E. Schmidt, a vice president at Sun. "Everyone says, 'Stop arguing! I just want it to be the same,'" Schmidt says. Others, including Kenneth H. Olsen, president of Digital, argue that the debate over operating systems constitutes a red herring. The operating system "is probably way down on a list of what's most important to standardize," Olsen said in a recent speech. "Let's get everyone to agree on other levels."

Among those other levels is the user interface, which consists of code for porting applications to the rest of the system and the "look and feel" that a person sees on screen. Here, once again, there is a rift in the industry. OSF has adopted an approach called Motif; UNIX International supports another known as Open Look but says it will consider additional interfaces if they meet specific criteria.

Between the operating system and the user interface are several layers of standards that are being widely embraced. POSIX 1003.1, developed by the IEEE, specifies a collection of internal services, such as how applications should retrieve and read files. About a dozen other POSIX standards are being developed. And most manufacturers now employ a technique called X Window for splitting a screen into separate viewing areas, or windows.

On the other hand, subtle differences among systems that basically conform to the standards can still cause havoc for users. Randall L. Frank, director of information technology at the University of Michigan, points out that there are more than half a dozen major different microprocessor chips for workstations. Even if workstations use the same operating system, subtle differences in the mechanics of their chips may force engineers to spend hours trying to port an application.

"Ninety-eight percent compatibility is no compatibility," Schmidt says. Yet in theory, open-system standards should be broad enough to work with many different chips, says Jim Isaak, who directs the IEEE's POSIX efforts. The real concern, he adds, is that the standards will be inadequate or not ready in a timely fashion.

To hardware makers, the road to open systems looks harrowing. When different architectures can run the same software, the machines become commodities with low profit margins, observes Franco Agostinucci, a vice president at Olivetti. Vendors will have to work hard to woo more customers to make up for lost profits.

Some companies are still trying to dodge standards, Agostinucci adds; they may claim to support open systems, but they continue to push their own technology. Isaak wonders: "Will the major vendors provide the same level of support and encouragement for their products that meet the open-systems criteria as they do for other approaches?" Until the standards are sufficiently complete, he says, manufacturers are likely to continue pouring most of their energy into supporting their own technology.

Open systems will consequently raise the level of competition in the computer industry, Schmidt predicts. This makes the business of picking standards enormously important and risky. "Innovation [will become] so quick that you have to hope your work becomes a standard," Frank observes. Otherwise it might be eclipsed by other technology. "If we can find a common ground," Schmidt says, "then we can get on with competing at a higher level"—namely on the issues of price, delivery, quality and performance. —Elizabeth Corcoran

Rational Drugs

Transforming drug research from an art into a science

In 1970 pharmaceutical researchers at Squibb were in a quandary. Through careful trial and error, they had developed a drug that had potential for combating hypertension. The prospects for commercializing it looked dim, however: derived from the venom of a Brazilian pit viper, the drug was prohibitively expensive to produce and had to be administered intravenously.

The researchers knew the drug worked by inhibiting the action of angiotensin converting enzyme (ACE),

an essential factor in the biochemistry of high blood pressure. Then they had an insight: Why not try designing a synthetic molecule to block the enzyme's action? Scientists had long dreamed of such an approach. Because ACE has such a highly specific architecture, the Squibb investigators scored a success: captopril, the drug they synthesized in 1975, became the first in a family of ACE inhibitors and the first "rationally designed" drug.

So far, ACE inhibitors are the most prominent rationally designed drugs to have reached the market. But aided by computer-modeling technology, which enables researchers to visualize the atomic structure of the molecules of interest, companies have begun exploiting rational design to speed the development of new drugs and to cut costs. "We see rational drug design as the key to the future," says Brian W. Metcalf, a vice president at SmithKline Beecham. "We are close to exhausting the traditional approaches."

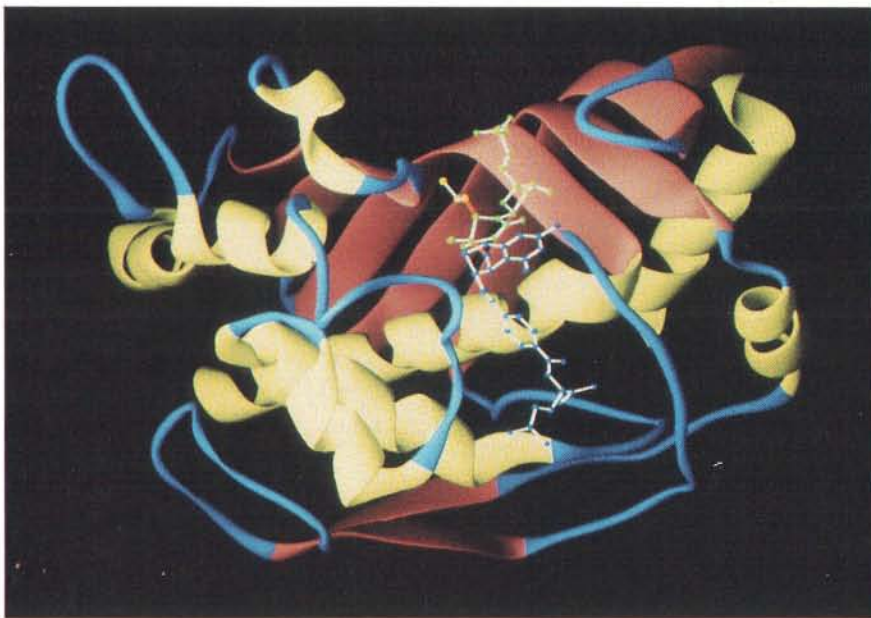
Pharmaceutical workers are quick to point out that the traditional, trial-and-error approaches were not irrational but simply dependent on luck. In contrast, biochemists employing rational design techniques try to engineer specific molecular structures to, say, block the action of an enzyme.

Squibb, which merged in October, 1989, with Bristol-Myers, is now testing its next generation of hypertension drugs in humans. Upjohn, Abbott and Merck are testing inhibitors that

clip the enzymatic cascade that elevates blood pressure at a slightly earlier stage than the one at which ACE inhibitors work. Abbott is also developing a therapeutic agent for Parkinson's disease; Merck is working on a drug for prostate disorders. Both SmithKline Beecham and Genentech are working on rationally designed chemotherapies for AIDS.

Start-up companies, too, see an important role for rational drug design. Agouron Pharmaceuticals in La Jolla, Calif., which has focused exclusively on rational drug design since it was founded in 1984, is working on an inhibitor for an enzyme essential to the frenzied proliferation of cancerous cells. BioCryst in Birmingham, Ala., hopes to improve the effectiveness of conventional chemotherapeutic drugs by blocking enzymes that attack the agents on their way to tumors. And Vertex Pharmaceuticals in Cambridge, Mass., founded barely a year ago, is trying to design suppressants to treat autoimmune disorders such as lupus, rheumatoid arthritis and diabetes.

Pharmaceutical companies hoping to exploit rational drug design still face formidable challenges. There is a shortage of researchers trained in the techniques. Moreover, before designing an inhibitor, investigators must be able to visualize the three-dimensional structure of the protein they are targeting; to do so, they typically rely on crystallized proteins that can be examined by X-ray crystallography.



COMPUTER MODEL OF THREE-DIMENSIONAL STRUCTURE of an enzyme, thymidylate synthase (ribbonlike structure), enables investigators at Agouron Pharmaceuticals to design an inhibitor (stick-figure structure) that will bind tightly to the enzyme and block its action. Software developed by M. Carson.

Yet growing well-formed crystals is tricky. "There is no 'love potion number nine,'" laments Alexander McPherson, a pioneer in the field and founder of Cryschem in Riverside, Calif.

By churning out large quantities of proteins, recombinant technologies have improved the odds of obtaining usable crystals, notes Peter Johnson, chief executive of Agouron. And according to Charles E. Bugg of the University of Alabama at Birmingham, "space grows crystals better than the best ever grown on the earth." In a recent issue of *Science*, Bugg, McPherson and 22 collaborators reported good success in growing crystals on board the space shuttle *Discovery* in September, 1988. About 50 groups are waiting for a chance to send their proteins into space. A few firms, including BioCryst, even hope to start up service companies that would grow crystals in space for others.

Rational drug design will not replace all other traditional drug discovery methods, notes Yvonne C. Martin, a project leader at Abbott. Instead it "changes at a fundamental level how people are practicing the art" by requiring researchers with more specialized knowledge. And although such design methods may produce a likely drug candidate more quickly, the longest part of the drug development cycle—namely, human clinical trials—remains unchanged. —Deborah Erickson

Tensaiji

Whiz kid wins business—even in Japan

"The nail that sticks out gets hammered down."

—Japanese saying

In a nation of conservative businessmen, Kazuhiko Nishi is a nail that resists hammering. He dropped out of the prestigious Waseda University to start his own company; before he was 30 Nishi gained a reputation as one of Japan's prophets of the personal-computer age. Now at 33 he can boast another accomplishment. This past September his company, the ASCII Corporation, became the first maker of personal-computer software to float its stock on Tokyo's over-the-counter market.

The U.S. tends to regard the talent and energy of its many young entrepreneurs as a unique asset. But Japan enjoys a similar endowment. It consists of young men who, like Nishi, are the kindred spirits—and future com-



COMPUTER ENTREPRENEUR Kazuhiko Nishi founded the ASCII Corporation about 13 years ago; among ASCII's many software packages is an operating system called MSX for personal computers, based on Microsoft's MS-DOS.

petitors—of the entrepreneurial hackers who exemplify the garage-to-riches spirits of California's Silicon Valley.

As a teenager Nishi was captivated by the potential he saw in early personal computers. In 1977 he and two friends chipped in \$10,000 each to start ASCII and to publish a PC newsletter. A year later Nishi flew to the U.S. and met another entrepreneur, William Gates, whose fledgling company, called Microsoft, was selling PC software. Nishi persuaded Gates to let ASCII sell the software in Japan.

The young Japanese businessman rapidly carried off some remarkable successes. He convinced the giant NEC Corporation to build Japan's first personal computer with the assistance of Microsoft. Nishi later talked Kyocera, a large ceramics company, into manufacturing a lap-top computer designed by Gates and himself. As Microsoft became a household word in the U.S., ASCII grew to become the larg-

est supplier of PC software in Japan.

Nishi and Gates split up in 1986 after a series of disagreements, according to published accounts. Among them: Nishi thought software vendors should design special-purpose microprocessors to control aspects of PC operation instead of relying on operating-system software. Nishi consequently broadened ASCII's business from magazine publishing and software development to include large-scale integrated-circuit design and communications networking.

"My expertise, our focus, is on applied technology—that is, we catch the wave of basic technology [such as large-scale integration] and conduct our R&D with a product in mind," Nishi says. Because ASCII contracts out manufacturing, he observes, "we can talk to everybody. All Japanese companies are our clients, as are many overseas firms. They all give us suggestions."

Sometimes those suggestions are veiled requirements: Zenith Data System's lap-top personal computer, the MinisPort, included an ASCII-designed video-controller chip. But the ASCII prototype was rejected by Zenith; the chip had to be redesigned to make it compatible with IBM products. "Many of our [sets of] chips that failed to sell did so because we stopped developing them after initial rejection by the customer," Nishi acknowledges, but "in all the cases where we listened to our clients and redesigned the chips, we were successful."

Even after the breakup with Microsoft, ASCII has grown briskly. Revenue rose about 20 percent in fiscal 1989 to approximately \$165 million (roughly one fifth of the revenue grossed by Microsoft). Profits have lagged, though: ASCII recorded just over \$2 million in net income in 1989, largely because of the impact of investments made as ASCII diversified, according to analysts in Tokyo.

Observers say Nishi bears careful watching. After all, as Peter G. Wolff of Kidder, Peabody & Company in Tokyo points out, Kazuhiko Nishi "rose out of the maelstrom of hundreds of software companies here, and he did it his way." —Stuart M. Dambrot

THE ANALYTICAL ECONOMIST

Tweaking the aggregates

"In God We Trust" is imprinted on all U.S. currency. In fact, the country places its economic faith in the seven governors of the Federal Reserve. Since fears of recession re-

placed fears of inflation last summer, stock-market analysts and the rest of the nation have hung on auguries of the Fed's actions. How do these seven people (along with the five who join

them on the Federal Open Market Committee) choose whether to push the economy up or down? What tools do they have to effect their decisions?

The Fed seeks stable prices. Chairman Alan Greenspan and his fellow governors have two primary instruments for achieving this goal: changing interest rates or manipulating the money supply (the ready cash flowing through the economy). The Fed can peg these monetary variables wherever it wants by buying or selling government bonds and by changing the discount rate (the rate the Fed charges banks for short-term loans).

The Fed's actions are felt first in financial circles, but tightening or easing of the money supply eventually affects the entire economy. Former governor Andrew F. Brimmer explains it this way: the Federal Reserve varies the rate at which it supplies money to the banking system. Faced with additional cash on hand, commercial banks try to put the extra money to work by lowering interest rates, which makes loans cheaper and easier to get. The lower price for money increases demand; consumers and businesses buy more goods and equipment, which stimulates the economy. If existing capacity can satisfy the new demand, the economy grows—if not, the increased money supply and lower interest rates trigger inflation. Conversely, reductions in the money supply slow the economy.

Controlling the money supply and controlling interest rates would seem to be the same; it is in their mechanisms and subtle side effects that they differ. When the Fed depresses the money supply by selling bonds (thereby taking cash out of circulation), bond prices drop in response to the extra supply, thus raising their effective yield and finally pushing up interest rates in general. A rise in the discount rate (which raises interest rates directly) encourages people to exchange cash for bonds, thus eventually depressing the money supply. The instrument individual governors prefer strongly reflects their allegiance to one or another economic school.

Monetarists, for instance, generally believe that the money supply directly controls the economy by limiting the funds available for investment and consumption. Keynesians, in contrast, claim that interest rates shape economic activity by encouraging businesses and consumers to borrow money to invest in new equipment and to buy goods.

There is little agreement on which camp prevails. The Fed tends to base

its monetarist decisions on Keynesian theory, says David E. Lindsey, deputy director of the Fed's division of monetary affairs: even if the governors believe interest rates are most important, picking the right level is difficult; it might be better for the economy if the Fed controlled the money supply and let the market fine-tune interest. Yet according to former governor H. Robert Heller, "you steer by interest" over the short term. Money-supply growth is the crucial variable only over periods of a year or more.

From early 1979 to late 1982 the Federal Reserve put aside Keynesian principles and took a strictly monetarist line. It controlled the money supply directly by buying and selling bonds and letting interest rates float as high as necessary to rein in money-supply growth. Before and since then the board has taken a more eclectic approach. "We follow Milton Friedman [the godfather of monetarism]," a Fed spokesman comments, "but not out the window."

This eclectic approach means that the Fed governors base their hunches about inflationary pressures—and the actions required to stifle them—on all kinds of economic indicators, from industrial capacity to housing starts. The indicators they monitor, Heller says, reflect the fact that inflation is a sequential process: it shows up first in commodities, then in prices for intermediate goods, then in consumer prices. By the time wages begin rising, it is too late for the Fed to act.

Long-term interest rates, yet another portent, reflect expectations of inflation a year or more ahead, Heller notes. Long-term rates rising as short-term rates fall could be a sign that the economy is about to become overheated. Such a warning would bring decisive action; although the Fed monitors many economic signals, it acts to curb inflation. Unemployment or the health of particular companies is not within the Fed's purview, Lindsey says: central bankers neither have nor want the tools for industrial policy.

The Fed must also contend with the fact that its decisions have only delayed effects on the U.S. economy. The actions of the Fed during November of 1989, for example, will not be felt throughout the economy until about March through May of 1990, governor Wayne D. Angell says. Worse yet, he laments, November's decisions were based mostly on data about what the economy was doing in October or even July and August. To mitigate this problem, Angell focuses on commodity prices, "which provide a daily measure

of inflationary pressures and tend to lead the behavior of the rest of the economy, he says.

As if dealing with such leads and lags were not bad enough, the Fed must rely on data virtually all of which are flawed in some way. Once upon a time, the data on the money supply included only cash in circulation and immediately available bank deposits; those were the sole sources of funds for immediate investment and consumption. As financial deregulation allowed interest-bearing checking accounts, however, the line between investments and cash on hand blurred. Today consumers can write checks against money-market funds or even use credit cards to borrow instantly against the value of their homes. The Fed has revised its measures of the money supply, but the notion of what funds are available for economic activity has become ever more nebulous. The "money supply" that the Fed tracks and attempts to control now grows or shrinks depending on interest rates rather than on how people intend to use it for investment and consumption.

Federal Reserve governors, and the financial analysts who watch them, have searched for years for a "stable anchor" that might predict inflation and thus permit the Fed to control it. Such an anchor is hard to come by, in part because Wall Street's anticipations of Fed actions affect the market and the economy as strongly as do actual changes in policy. If the Fed lowers interest rates, Heller says, the market may anticipate further easing; a failure to push rates lower may have the same effect as raising them. The Fed's credibility in pursuing a policy is an important factor in making that policy work, Heller says.

Although Fed governors believe they have done reasonably well over much of the past decade, Angell cautions against complacency. Any successful method for regulating the economy carries the seeds of its own downfall as the market learns the technique and counteracts its effects. Keynesian economics worked well until people noticed that it predicted prices would always go up; monetarism quashed that expectation. Then, as the Federal Reserve held down growth during the 1980's, high demand for money "made monetarists look like fools," Angell comments. Now a more eclectic policy seems to be working, but the market will eventually decipher it as well, and the Fed will have to scramble again to stay in control.

—Paul Wallich and Elizabeth Corcoran

The Handedness of the Universe

From atoms to human beings, nature is asymmetric with respect to chirality, or left- and right-handedness. Clues are beginning to emerge that connect chirality on different levels

by Roger A. Hegstrom and Dilip K. Kondepudi

In 1848 Louis Pasteur, examining a certain salt of tartaric acid under a microscope, noticed that it formed two types of crystals, each one a mirror image of the other. He separated the two, dissolved each in water to form two solutions and shined a light beam through each. To his great surprise, one solution rotated polarized light clockwise, the other counterclockwise.

This beautiful discovery, which he made at the age of 25, led Pasteur to develop a theory of molecular structure. Little was known then about the structure of matter on such a small scale; Pasteur postulated that the two distinct shapes of the salt crystals and their ability to rotate light differently were derived from the fact that the molecules making up the salt were themselves of two types, one "right-handed" and the other "left-handed."

His research along these lines paved the way for another remarkable discovery in 1857. One day Pasteur found that molds had grown in a dish containing an optically inactive solution, one that did not rotate light. Instead of simply throwing away the "contaminated" solution—the common practice—Pasteur checked its effect on a light beam. The contaminated solu-

tion rotated light! Microorganisms had changed an optically inactive solution to an optically active one.

On the basis of his molecular theory, Pasteur reasoned that the original solution was optically inactive because it contained equal numbers of right- and left-handed molecules. The molds had reacted chemically with only one type, leaving the solution with a relatively large amount of the other. The imbalance made the solution optically active.

Thus, Pasteur realized that the chemistry of life has a preferred handedness. He came to view handedness as one of the clearest distinctions between living and dead matter and ultimately proclaimed it to be a profound fact of nature that went far beyond the chemistry of life. "Life as manifested to us," Pasteur wrote, "is a function of the asymmetry of the universe and of the consequences of this fact." Later, before the French Academy of Sciences, he made the grand conjecture, "*L'univers est dissymétrique.*"

Pasteur's conjecture turned out to be true to an extent that no one, perhaps even he, imagined. Modern science has revealed that mirror symmetry is often absent in nature: the universe is *dissymétrique* at all levels, from the subatomic to the macroscopic. Many questions about how this asymmetry arises remain unanswered, but in the past few decades some understanding has been gained as to how handedness at one level may give rise to handedness at another. In order to describe what is known and what is not, it is convenient to begin at the scale of everyday objects.

Chiral Asymmetry

Most objects found in nature are not identical with their mirror images and therefore are said to possess chirality, or handedness. To distinguish the two forms, they are often designated right-handed or left-handed. In the case of

some familiar chiral entities—hands or screws, for instance—the meaning of right- and left-handed is clear, but for objects such as an elm tree with many branches or for generally irregularly shaped things, the designation is somewhat arbitrary. When certain very simple objects such as spheres or triangles are reflected in a mirror, the resulting image is indistinguishable from the original object. Objects that are identical to their mirror images are termed achiral.

Not only objects but also processes, such as chemical reactions, may exhibit chirality. Certain atomic and nuclear interactions, for instance, display a preference for left or right. If all processes were chirally symmetric, one would observe in the real world an equal number of mirror-image systems displaying opposite preferences. That we do not is evidence that some processes in nature are asymmetric.

Although a chiral object and its mirror image are obviously different, there is no a priori reason that one should be superior to the other. Yet the real world usually does display a preference for one kind of chirality over another. This is strikingly demonstrated in the case of living organisms. Human beings, for instance, are structurally chiral: the heart is to the left of center, the liver to the right. People also display functional chirality. For example, although there is no apparent intrinsic advantage to either the left or the right hand, few people are ambidextrous. Why do individuals generally prefer one hand over the other? Many reasons can be postulated, but the correct one probably is not yet known.

Given that humans generally are not ambidextrous, the next question is: Why are most people right-handed? The dominance of the right hand over the left is universal, independent of race and culture. There would be no apparent disadvantages if most people were left-handed. The greater

ROGER A. HEGSTROM and DILIP K. KONDEPUDI work in the department of chemistry at Wake Forest University. Hegstrom received his B.A. from St. Olaf College in 1963 and his Ph.D. from Harvard University in 1968. His theoretical research has concentrated on the effects of electromagnetic and weak interactions in atoms and molecules. Kondepudi received a B.Sc. from the University of Madras in 1971, an M.Sc. from the Indian Institute of Technology, Bombay, in 1973 and a Ph.D. from the Center for Statistical Mechanics at the University of Texas at Austin in 1979. His interest in the origin of order and complexity has taken him through the disciplines of physics, chemistry and biology.

number of right-handed people seems to be just an accident. One might also ask why right- and left-handed persons are not born in equal numbers. Again, the answer is not known with any certainty, although it is plausible to argue that handedness is an inherited trait: once right-handedness became dominant, for whatever reason, it remained so.

There are other, less prominent examples of chiral asymmetry in organisms. Helical seashells spiral either like a right-handed screw or like a left-handed screw. Right-handed, or dextral, shells dominate—on both sides of the Equator. Among these right-dominated animals, left-handed individuals exist only as a result of mutations, which appear with a frequency ranging from about one in hundreds to one in millions, depending on the species. This "right-hand rule" is not universal, however: certain species—for example, the lightning whelk of the Atlantic coast—are predominantly left-handed, or sinistral. In rare instances left- and right-handed individuals occur in a species in almost equal numbers; the Cuban tree snail, *Liguus poeyanus*, is an example.

Like animals, most types of plants exhibit a preferred chirality. Bindweed

winds as a right-handed helix, whereas honeysuckle grows as a left-handed helix. Helical structure in organisms also has been found on the smaller scale of bacteria. Since the 1970's Neil H. Mendelson and his co-workers at the University of Arizona have investigated the bacterium *Bacillus subtilis*, which usually forms right-handed spiral colonies. Remarkably, as the temperature increases, the spiral becomes left-handed!

Chirality in Molecules

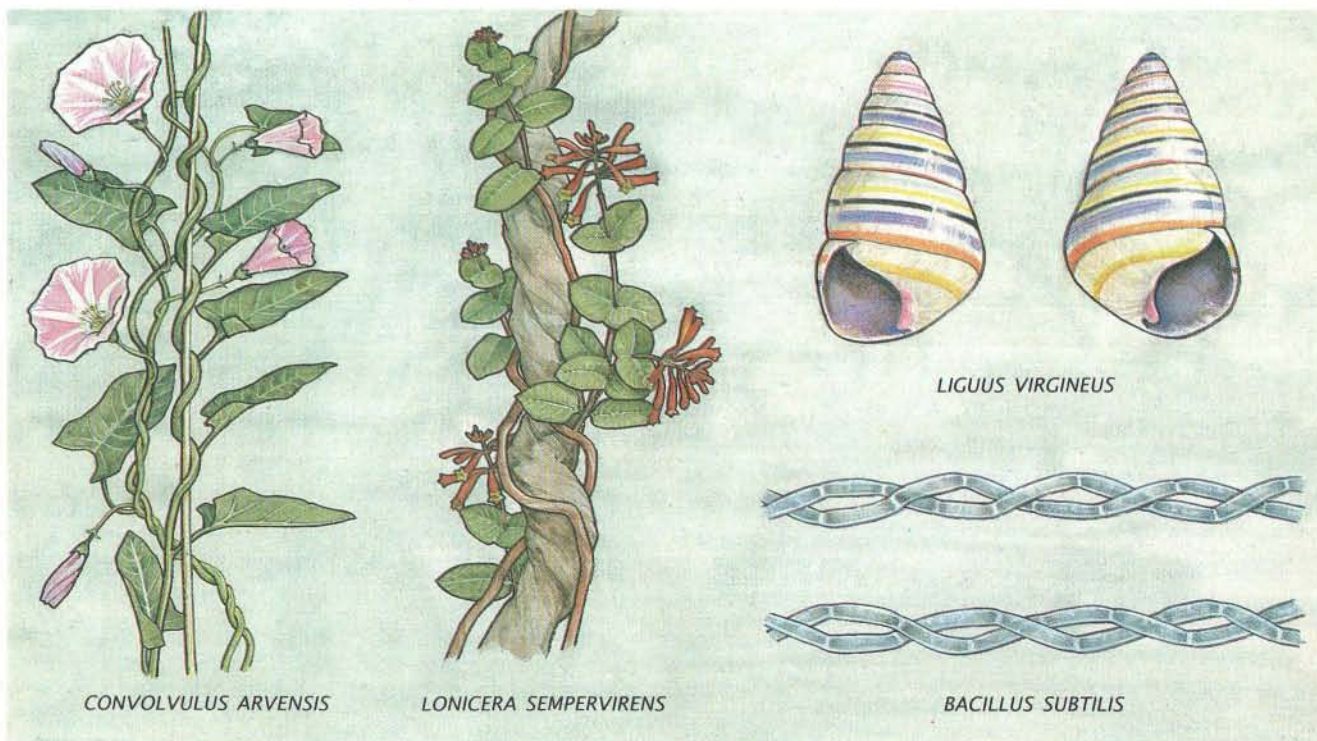
As Pasteur found, molecules can also be chiral. Chemists refer to mirror-image molecules as L-enantiomers and D-enantiomers; L and D stand for levo (left) and dextro (right), a relic from Pasteur's studies of optical rotation of light. Enantiomeric forms are found in many organic and inorganic substances and in essentially all molecules crucial for the development of life—specifically proteins, which are responsible for the structure and chemical regulation of living cells, and DNA, the molecule that carries genetic information.

A protein molecule is a polymer, that is, a long chain of smaller molecules—in this case, a chain of ami-

no acids. Although several hundred amino acids exist, all proteins are made from the same 20 amino acids. All the amino acids but one (glycine) are chiral, having L- and D-enantiomers. Strangely enough, proteins are made exclusively of L-amino acids. (In very rare cases short strings of amino acids—polypeptides—that contain D-amino acids serve a specialized biological role.)

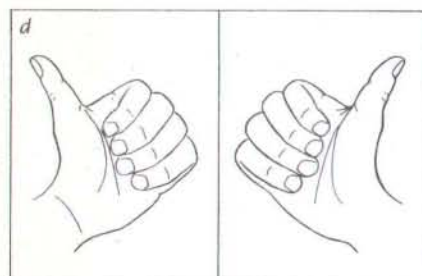
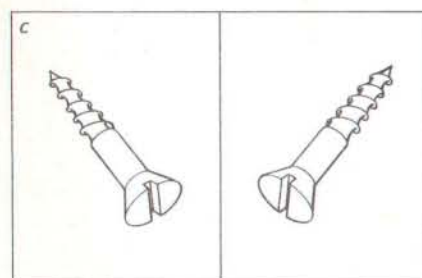
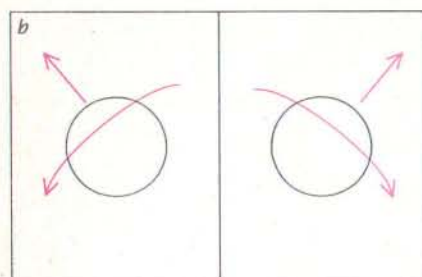
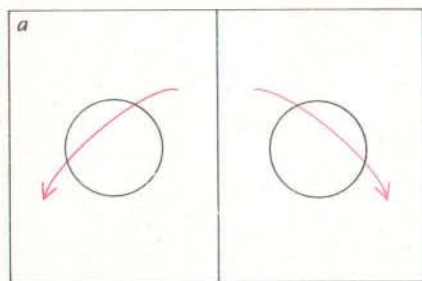
The main function of certain proteins, called enzymes, is to catalyze biomolecular reactions, including the synthesis of other proteins. The catalytic ability of enzymes depends crucially on their three-dimensional structure, which in turn depends on their L-amino acid sequence. Synthetic chains of amino acids made of both L- and D-enantiomers do not twist in the way necessary for efficient catalytic activity; they cannot form the regular winding structure, called the alpha helix, that is present in most enzymes.

Because of the chirality of its key molecules, human chemistry is highly sensitive to enantiomeric differences. An extreme example came to light in 1963 when horrible birth defects were induced by thalidomide. The defects were caused by the fact that whereas one enantiomer of this chiral com-



PREFERRED HANDEDNESS is a common trait of living things. Trumpet honeysuckle, *Lonicera sempervirens*, winds to the left; bindweed, such as *Convolvulus arvensis*, winds to the right—like the majority of helical plants. Snails, such as *Liguus virgineus*, are generally right-handed, but within a species, left-hand-

ed versions appear owing to mutations. The bacterium *Bacillus subtilis* normally forms right-handed spiral colonies; when heated, these change to left-handed. Atoms and molecules also are asymmetric with respect to left and right, but this has not yet been plausibly linked to the handedness of living objects.



CHIRALITY manifests itself in the distinction between left and right. Objects that cannot be superposed on their mirror images are termed chiral. A stationary sphere is identical with its mirror image and is said to be achiral; even if a sphere is spinning (a), its mirror image can be superposed on the original object by turning it upside down, and so a spinning sphere is also achiral. If the sphere is moving along its spin axis (b), the mirror image cannot be superposed on the original, and the object becomes chiral. By convention, if a spinning object behaves like a right-handed screw as it moves, it is termed right-handed; if it behaves like a left-handed screw, it is termed left-handed (c). The direction of spin is defined by the "right-hand rule": curl the fingers of the right hand in the sense of rotation; the thumb points in the direction of the spin axis (d). (Hands and screws are chiral objects and cannot be superposed on their mirror images.)

pound cured morning sickness, the other caused birth defects. Today the pharmaceutical industry pays careful attention to the separation of enantiomers. A less morbid case of enantiomeric sensitivity involves limonene, a compound found in lemons, oranges and perfumes. Here one can smell the difference: one enantiomer smells like lemons, the other like oranges.

Like proteins, the nucleic acids DNA and RNA are polymers that exist in nature in only one chirality. Each is composed of four types of subunits, each of which incorporates a chiral sugar group. Only the D-enantiomer of the sugar is present in nucleic acids. DNA and RNA ordinarily form right-handed helices as a result of the exclusive presence of D-sugars. The proper replication of nucleic acids depends on the activity of proteins made of L-amino acids, and so the relative chiralities of proteins and nucleic acids are intimately connected.

The great preference in the chemistry of life for L-amino acids and D-sugars over their mirror-image counterparts is peculiar for two reasons. First, except for extremely minor differences to be discussed later, the chemical properties of the L- and D-enantiomers are essentially mirror-symmetric. Second, when chiral molecules are synthesized in the laboratory from achiral building blocks, equal amounts of L- and D-enantiomers are produced unless painstaking care is taken to introduce an asymmetric agent during the synthesis.

There is a fundamental underlying reason for this symmetry: chemical reactions are essentially a result of the electromagnetic interaction of atoms. The electromagnetic force behaves in such a way that if a given process takes place, the mirror image of that process occurs with equal probability. Any force that gives rise to both a process and its mirror image with equal probability is termed parity-conserving. Because the electromagnetic force conserves parity, one would expect equal numbers of L- and D-enantiomers to inhabit the world. Why is this not so? We shall return to this question after examining chirality at the subnuclear scale.

Four Forces

All known elementary particles interact with one another through four types of forces: gravity, the electromagnetic force (responsible for ordinary chemical reactions), the strong nuclear force (which holds atomic nuclei together) and the less well-known

weak nuclear force. Until 1957 it was thought that nature was chirally symmetric at the scale of elementary particles—that is, that the four forces were parity-conserving. In that year it was discovered that the weak nuclear force does not conserve parity.

As its name implies, the weak force is relatively feeble, about 1,000 times less powerful than the electromagnetic force and 100,000 times less powerful than the strong nuclear force. The most familiar effect governed by the weak force is the production of beta rays in radioactive decay. Beta rays are actually energetic electrons and their antimatter twins, positrons. These particles have an intrinsic spin and hence, when they are moving along or against their spin axes, can be classified as either left- or right-handed. The surprising and now famous 1957 discovery of parity violation by Chien-Shiung Wu and her colleagues at Columbia University led to the recognition that beta particles emitted from radioactive nuclei have a definite chiral asymmetry: left-handed electrons far outnumber right-handed ones.

Further investigations of beta decay led to the discovery of the neutrino and antineutrino, electrically neutral particles that are also emitted in beta decay and that always travel at the speed of light. Like the electron, the antineutrino emitted by radioactive matter has a spin; unlike the electron, it exists only in the right-handed form. No one knows why chiral asymmetry exists at such a fundamental level. Radioactive antimatter emits an excess of right-handed positrons (antielectrons) and only left-handed neutrinos. Right-handed neutrinos and left-handed antineutrinos seem not to exist in the universe.

For the next decade or so it was believed that parity nonconservation was confined to nuclear reactions. Phenomena such as chemical reactions or interactions between atoms and light, which depend on the electromagnetic force, appeared to conserve parity. In the late 1960's, however, Steven Weinberg, now at the University of Texas at Austin, Abdus Salam of the International Centre for Theoretical Physics in Trieste and Sheldon L. Glashow of Harvard University developed a theory that unified the weak and electromagnetic forces [see "Unified Theories of Elementary-Particle Interaction," by Steven Weinberg; *SCIENTIFIC AMERICAN*, July, 1974]. Their theory predicted a new "electroweak" force between an atom's electrons and the protons and neutrons in its nucleus. The existence of this force, which

does not conserve parity, was confirmed in the 1970's.

Because the electroweak force distinguishes between left and right, atoms and molecules that were previously thought to be achiral must be chiral in some way. Furthermore, enantiomers such as L- and D-amino acids must differ with regard to physical properties, such as energy, that depend on their handedness.

It is now evident that the world is chirally asymmetric at all scales, from the scale of elementary particles upward. How do the asymmetries arise? Are chiral symmetries at one level linked to those at another, or are they independent? We shall attempt to answer these questions, insofar as it is possible to answer them, beginning at the scale of elementary particles.

Chirality of Elementary Particles

At rest, an elementary particle such as an electron or a positron is spherically symmetric and hence achiral. But if a spinning particle is moving in either direction along its spin axis, it becomes chiral. If it behaves like a right-handed screw as it moves, it is said to be right-handed; if it behaves like a left-handed screw, it is said to be left-handed.

Chiral asymmetry at the subatomic level is fundamentally connected to parity nonconservation. According to the Standard Model of elementary particles propounded by Weinberg, Salam and Glashow, the electroweak force distinguishes between left and right through "weak charged currents" and "weak neutral currents." The strength of these currents—referred to as the *W* and *Z* forces—between any two elementary particles depends on the distance between the particles and on their "charges."







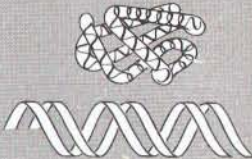
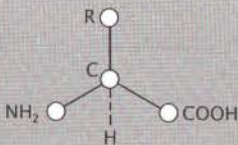
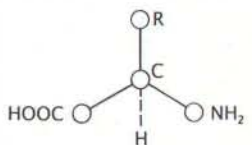

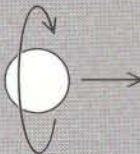
We use the term charge here in analogy to electricity. The electron has a negative electric charge, and the electrical force between any two electrons is repulsive. In contrast, the weak *W* charge is nonzero for a left-handed electron but zero for a right-handed one. Therefore, a right-handed electron simply does not "feel" the *W* force. This is considered a fundamental property of the weak force; at present there is no deeper understanding of it. One result of this asymmetry is that nuclear beta decay, which is governed by the *W* force, produces mostly left-handed electrons.

As for the *Z* force, left- and right-handed electrons have *Z* charges of opposite signs and approximately equal magnitudes. The difference in sign

causes right-handed electrons to be attracted to the nucleus by the *Z* force and left-handed ones to be repelled. (These statements about the effects of the *W* and *Z* charges on chiral electrons are strictly valid only when the electrons are at high energies, traveling near the speed of light. The concepts are nonetheless useful for un-

derstanding the chiral asymmetries in low-energy electrons.)

In a looking-glass world, beta decay would produce right-handed electrons, and the *Z* force would attract left-handed electrons to the nucleus. These processes are not observed in the real world, however, which is another way of stating that the weak

HELICAL SEASHELLS		
HELICAL PLANTS		
HELICAL BACTERIA		
PROTEINS AND DNA	VERY RARE IN NATURE	
AMINO ACIDS		
CHIRAL CURRENTS IN ATOMS	NOT FOUND IN NATURE	
HELICAL NEUTRINO		NOT FOUND IN NATURE

PREFERENCE between left and right is displayed by nature on many levels. Colored boxes indicate the predominant handedness. Most helical seashells are right-handed, but some left-handed species and mutants exist. Winding plants are also predominantly right-handed. Helical bacteria come in right- and left-handed versions. Ordinarily, proteins and DNA wind in right-handed helices; left-handed versions are rare, and true mirror-image versions do not appear in nature. Right- and left-handed amino acid molecules exist at different energy levels as a result of the asymmetric weak nuclear force; those in organisms are almost always left-handed. The weak force also affects the way electrons orbit the nucleus and so causes atoms in general to become right-handed. The elementary particle known as the neutrino exists only as a left-handed object: its direction of spin points contrary to its direction of motion.

force is chirally asymmetric and that parity is not conserved.

Atoms and Molecules

An important consequence of the weak Z force between electrons and nuclei is that all atoms are chiral. Because of the Z force, when an electron is near the nucleus, its direction of motion is partially aligned with its spin axis, which makes it right-handed [see illustration below]. This means that the electron orbit, which would be circular in the absence of the Z force, becomes a right-handed helix in

the vicinity of the nucleus. Because the interaction that causes the helical electron motion does not conserve parity, the mirror-image atom with a left-handed helical electron flow does not exist in nature.

Given the extremely low strength of the weak force, one might expect this helical motion to be unmeasurable. For instance, the Standard Model predicts that in the most favorable experimental setup, light passing through an atomic gas should be rotated by a scant 10^{-5} degree—the angle subtended by a hand at a distance of roughly 1,000 kilometers. And yet during the

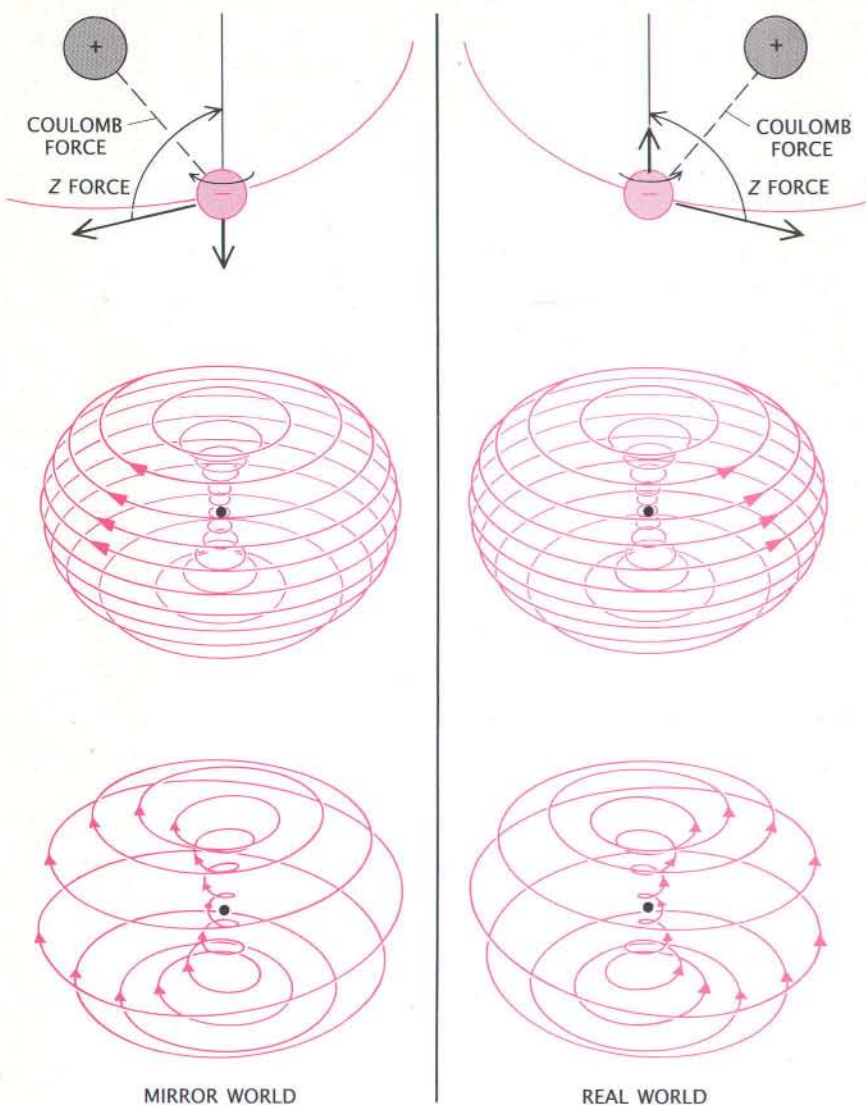
past decade experimental support for the chirality of atoms has been obtained, including the observation of rotations of the predicted amount [see "An Atomic Preference between Left and Right," by Marie-Anne Bouchiat and Lionel Pottier; *SCIENTIFIC AMERICAN*, June, 1984]. Here is one clear instance in which a chiral asymmetry at the level of elementary particles causes a chiral asymmetry at the higher level of atoms.

On a slightly larger scale, the Z force causes a chiral molecule to exist in a higher- or lower-energy state than that of its enantiomer. The split comes about in a subtle way. First, suppose that one models the chiral molecule as a helix, and imagine the Z force to be "turned off." If an electron with spin "up" is moving "up" the helix, it will be right-handed; if a spin-up electron is moving "down" the helix, it will be left-handed. Because probabilistically equal numbers of electrons in a molecule are moving up and down, one would expect the average electron chirality to be zero.

The ordinary parity-conserving electromagnetic forces between the electrons and the nuclei in the molecule, however, tend to align the axis of each electron's orbit against its axis of spin; this phenomenon is referred to as spin-orbit coupling. For a right-handed helical molecule, spin-orbit coupling favors down-spiraling for spin-up electrons and up-spiraling for spin-down electrons. In either case the spin axis of the electron tends to be aligned against the electron's direction of motion, so that in a molecule shaped as a right-handed helix, spin-orbit coupling produces predominantly left-handed electrons. In regions where the molecule is shaped as a left-handed helix, right-handed electrons predominate. As a result, molecules display regions of differing electron chirality [see illustration on opposite page].

Now switch on the Z force. Because the Z force interacts in different ways with right- and left-handed electrons, it produces an energy shift in the molecules: the energy of one enantiomer is increased and that of the other is decreased.

The Z force is so small that its effect on the chemical properties of molecules has not been observed. An interesting theoretical result, however, has been obtained by Stephen F. Mason and George E. Tranter of Kings College, London. Between 1983 and 1986 they performed detailed calculations of the energies of several L- and D-amino acids, taking into account the



ATOMS become chiral under the action of the weak nuclear Z force. At the top right an electron with spin "up" is shown in orbit around a nucleus; its mirror image is at the top left. Without the Z force, the paths of the electron flow would resemble those in the middle drawings. The nucleus is at the centroid of each atom. If the mirror image is flipped upside down, the new electron paths can be superposed on the original ones, and so these paths are achiral. With the Z force present, the direction of the electron's motion tends to align with the direction of its spin. The result is shown at the bottom right. The paths are now chiral: the electrons travel up along the inner, right-handed helix and down along the outer, left-handed helix. The mirror-image atom, shown at the bottom left, does not exist in the real world. For this drawing the effect of the Z force was magnified by a factor of 10^{10} .

asymmetric Z force. The expected energy split between the enantiomers emerged; curiously, in all cases the biologically dominant L-enantiomer was found to have the lower energy.

Basic principles of statistical mechanics require that in any equilibrium situation the lower-energy form should be more abundant than the higher-energy form. Mason and Tranter showed that L-amino acids should outnumber D-amino acids by one part in 10^{17} . Such an infinitesimal difference explains why L- and D-enantiomers are found in the laboratory in essentially equal numbers. Still, one cannot help but wonder whether this minute difference, caused by the weak nuclear force, is somehow connected with the dominance of L-amino acids and D-sugars.

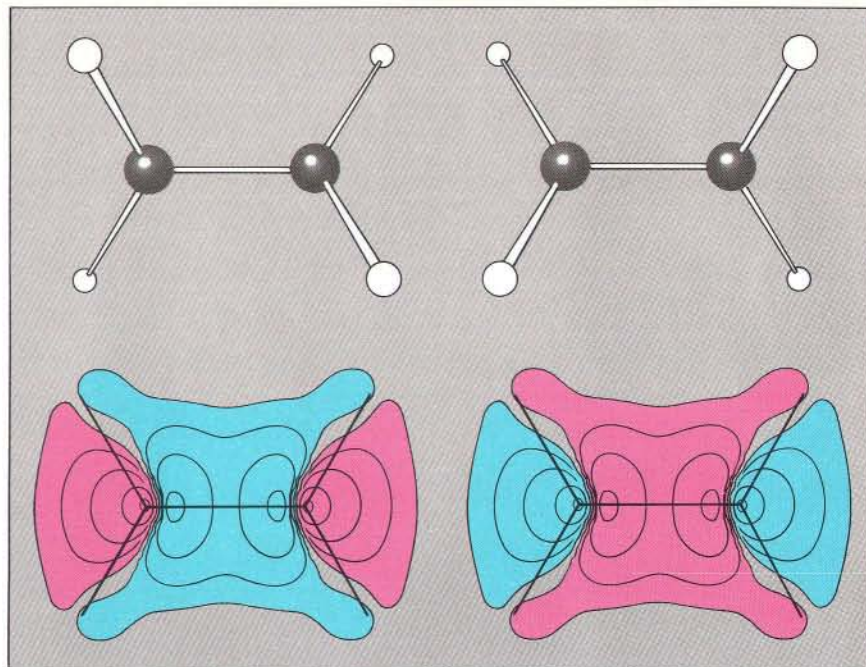
Chiral Symmetry in Life

So little is known about the origin of life that one cannot speculate about its causes with any confidence, but since the first experiments of Stanley L. Miller of the University of Chicago in the 1950's, scientists have developed a good picture of how a variety of biologically significant molecules could have arisen on the primitive earth. Somewhere in the course of the chemical evolution that led from atoms to life, the chiral asymmetry of biomolecules was established.

This raises three important questions. How could biomolecules with a chiral preference have arisen from chemical reactions that are identical for the two enantiomers? Is the dominance of L-amino acids and D-sugars over their mirror images in any way linked to the weak force? Was chiral asymmetry a precondition without which life could not have arisen, or did the asymmetry arise later—as a consequence of biological, rather than chemical, evolution? We shall address these questions one by one.

Paradoxical though it may seem, mirror-symmetric chemical reactions can produce unequal amounts of L- and D-amino acids through a phenomenon called spontaneous symmetry breaking. In this case, a symmetric state is one with equal numbers of L- and D-forms; the asymmetric state is one in which one form dominates. Spontaneous symmetry breaking is a mechanism by which a system "spontaneously" goes from a symmetric state to an asymmetric one.

Spontaneous symmetry breaking occurs only under specific physical conditions. It cannot occur in a system closed to the inflow of energy and



TWISTED ETHYLENE is a simple chiral molecule consisting of two carbon atoms and four hydrogen atoms (C_2H_4). The top drawings show the D- and L-enantiomers at the right and the left, respectively. In ethylene a phenomenon known as spin-orbit coupling, which tends to align an electron's spin against its orbital angular momentum, produces regions of differing electron chirality. The bottom drawings, based on calculations by one of the authors (Hegstrom) and his student Melinda S. Montgomery at Wake Forest University, show these regions as viewed from "above" the line connecting the two carbon atoms. Red shading indicates the regions where the electrons are right-handed, blue shading the regions where they are left-handed. Mirror reflection reverses the regions of chirality. The weak Z force acts in an opposite way on left- and right-handed electrons, so that the mirror reflections are subtly different: the L-enantiomer of ethylene has a lower energy than the D-enantiomer.

matter. Such a system will proceed toward thermodynamic equilibrium, a state in which the concentration of a molecule depends only on that molecule's energy and entropy. Because the energies of L- and D-enantiomers are equal (ignoring the tiny energy difference caused by the Z force), in this state the numbers of L- and D-enantiomers will be equal, and the state will be chirally symmetric.

If the system is open to the inflow of energy or matter, however, it is no longer in thermodynamic equilibrium. Spontaneous symmetry breaking then can become operative and can throw the system into a chirally asymmetric state, one that has unequal amounts of the enantiomers.

In 1953 Sir Frederick Charles Frank of the University of Bristol developed a simple model to illustrate how spontaneous symmetry breaking might operate in a chemical system consisting of two molecular species. Frank's model assumes that each species is capable of replication and that the presence of one diminishes the population growth rate of the other; that is, they compete. The replication rates

for the species are identical, as is each one's effect on the other. Nevertheless, as soon as one species becomes slightly more numerous than the other (for example, by means of a random statistical fluctuation), the more numerous species quickly becomes completely dominant. The symmetric balance between the two types of molecules is unstable and spontaneously evolves into an asymmetric state in which one type dominates.

It is easy to imagine how this would work on a biological level. Even if the mirror image of life as we know it once existed on the earth, competition between the two types might have resulted in the extinction of looking-glass life. Frank's model shows that this is also possible on the molecular scale, thereby demonstrating how an excess of L-amino acids or D-sugars could have arisen from a primordial soup in which both enantiomers were initially on an equal footing.

The Weak Force Again

We now turn to the second question: Is it possible that the weak nuclear

force is responsible for the dominance of L-amino acids and D-sugars? Ever since the discovery of parity violation, there have been attempts to invoke beta decay and related phenomena as mechanisms that could lead to an excess of one enantiomer. Frederic Vestner and Tilo L. V. Ulbricht, who were at Yale University in 1957 when parity violation was discovered, noted that beta electrons, because they are predominantly left-handed, emit predominantly left-handed electromagnetic radiation (radiation that is polarized and rotated to the left). Vestner and Ulbricht proposed that left-handed radiation decomposes one enantiomer preferentially, leaving a net excess of its mirror image. The expected asymmetry produced by the Vestner-Ulbricht process, however, is extremely small and has yet to be detected experimentally.

Beta particles can also decompose chiral molecules directly. One of us (Hegstrom) has calculated that the relative difference in the rates of such decomposition for L- and D-enantiomers is about one part in 10^{11} . Experiments by Arthur Rich, James C. Vanhouse and their co-workers at the University of Michigan have found that the difference is indeed less than one part in 10^9 .

Yet another candidate is the Z force itself, which can affect the production rates of L- and D-amino acids. As previously noted, however, the effect of the

Z force is so minuscule that the expected difference would be about one part in 10^{17} . For such a small asymmetry to have produced the observed dominance of L-amino acids and D-sugars, some amplification mechanism must have been operating.

One of us (Kondepudi) and George W. Nelson, now at the Los Alamos National Laboratory, have shown theoretically that such a mechanism indeed exists in nonequilibrium chemical systems. It is referred to as noise averaging by communications engineers, who exploit it to extract a signal from a noisy background. Imagine a pool of water in which two enantiomers compete with each other, as in Frank's model. Many random influences will tend to favor the survival first of one enantiomer and then of the other. These fluctuations are much larger than the effect of the weak force, but because they are random, they tend to cancel out. Given enough time, the small systematic effect of the weak force will influence the handedness of the symmetry breaking and push the system to a dominance of one enantiomer over the other.

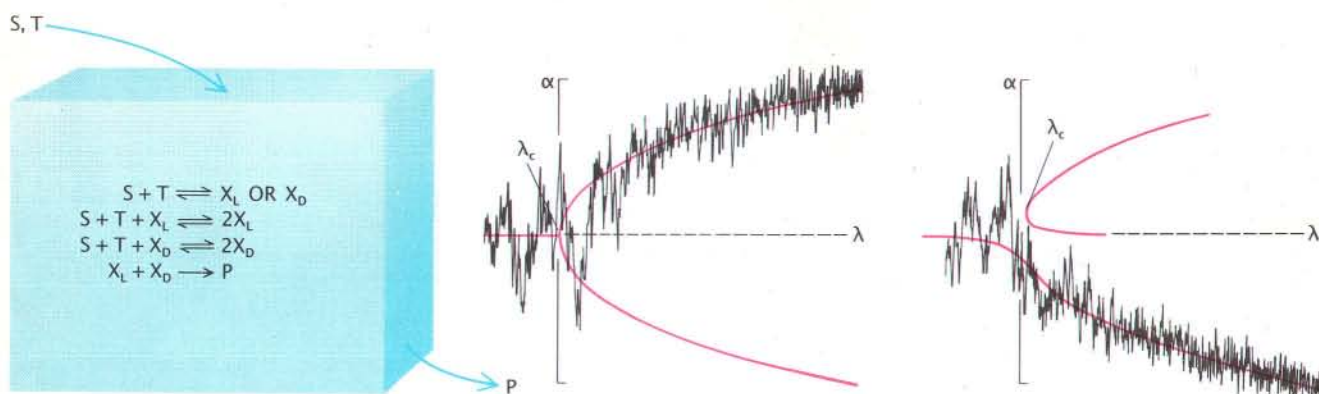
What conditions are necessary for the noise averaging to operate, and what time scales are involved? There should be a more or less constant flow into the pool of the achiral reactants needed to produce the enantiomers. The system will therefore be open and far from equilibrium, ensuring that

spontaneous symmetry breaking can take place. The reactants should produce enantiomers that replicate and compete with each other. And the pool should be large enough and sufficiently well mixed (over an area of about 10 square kilometers and a depth of several meters, roughly) to eliminate largely the net effect of random fluctuations. If these conditions are satisfied, the weak nuclear force should be capable, over a period of from 50,000 to 100,000 years, of strongly influencing the outcome of the symmetry-breaking process. After this time there is at least a 98 percent chance that nearly all the molecules—amino acids, in this instance—will be left-handed (assuming that the weak force favors L-enantiomers). In such an environment, chirally asymmetric life based on L-amino acids could evolve.

Such a slow chemical process is difficult to observe in the laboratory. An elegant electronic simulation by Frank E. Moss of the University of Missouri at St. Louis and Peter V. E. McClintock of the University of Lancaster has confirmed the existence of the predicted mechanism, but no such mechanism has yet been observed in a real chemical system.

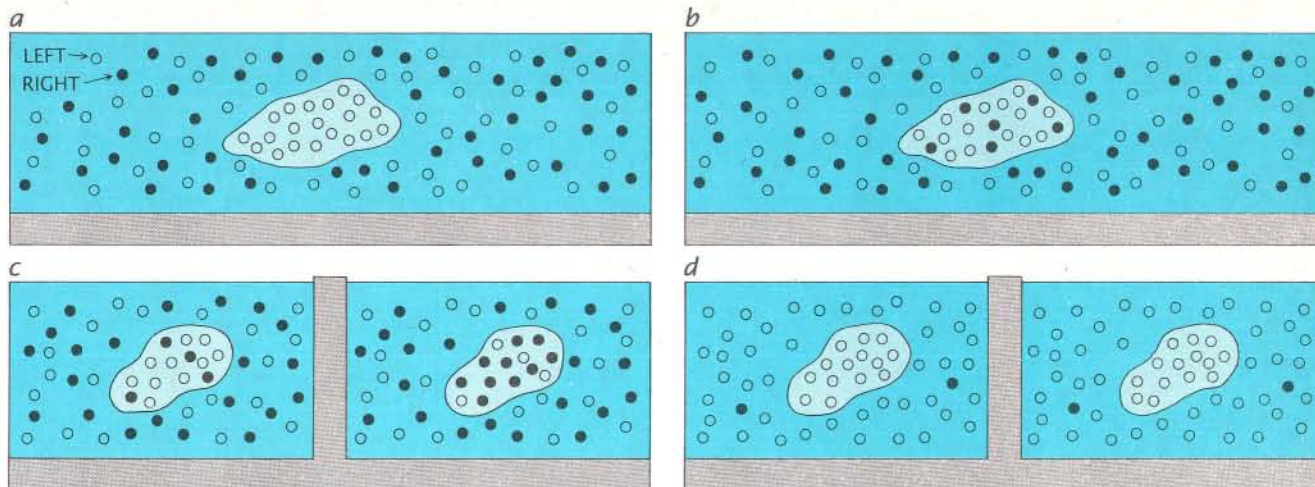
Before or after Life?

We have presented several models to show how chiral asymmetry might have arisen in biomolecules. The fi-



AUTOCATALYSIS AND SYMMETRY BREAKING are demonstrated in a simple chemical model. Two achiral molecules, S and T, are pumped into a pool of water (left). They react to form the chiral molecule X in either of its enantiomeric forms, X_L or X_D . X may react again with S and T to produce a second X_L or X_D ; this self-replication is termed autocatalysis. X_L and X_D may also annihilate each other by producing a product P. If none of these reactions favors the L- or the D-enantiomer, the concentrations of X_L and X_D should remain equal. The reaction-rate equations show, however, that the balance between autocatalysis and mutual annihilation is unstable. The critical parameter is λ , the product of the concentrations of S and T. When λ is increased past a critical value λ_c , the system will flip into a state where X_L or X_D is favored, although which state is

chosen is entirely random. The symmetry between L and D is "broken spontaneously." Alpha (α), the difference between the concentrations of X_L and X_D , is a measure of this asymmetry. Frank E. Moss of the University of Missouri at St. Louis and Peter V. E. McClintock of the University of Lancaster simulated this model electronically. They found (center) that as λ increased, symmetry was broken and X_D became dominant, although the dominance of X_L was equally likely. They also altered the simulation to give a small systematic advantage to X_L , analogous to the possible effect of the Z force (right). As λ increased, the system almost always followed the lower branch, where X_L dominates; the upper branch, where X_D dominates, became an improbable outcome. Such a model may explain the dominance of L-amino acids over D-amino acids in nature.



CHIRAL LIFE CHEMISTRY may be a relic of prebiotic conditions or an artifact of the life process. One theory holds that when life appeared in the primordial soup, the first cell formed containing all L-amino acids (a). This would have been extremely improbable, however, if the soup was composed of an equal mixture of left and right enantiomers. Another possibility is that the first cell randomly formed with a slight excess of L-amino acids (b), and evolutionary selection favored life based

on just one enantiomer. Some have proposed that life arose in many places simultaneously in both L- and D-amino acid-based forms (c); these forms competed, and life based on D-amino acids became extinct. An alternative view, investigated by the authors, is that spontaneous symmetry breaking produced near chiral homogeneity in each of the many places where life appeared (d). The parity-violating weak force influenced the symmetry-breaking process in favor of L-amino acids.

nal important question is whether this asymmetry arose before or after the appearance of the first primitive life, the "first cell." Based on current knowledge of the structure and function of biopolymers, it is difficult to understand how a protein or nucleic acid consisting of both L- and D-monomers could function. Experiments show that strings of amino acids containing both L- and D-acids do not correctly form the alpha helix shape that is crucial for the catalytic functions of proteins. Without homochirality (the situation in which all amino acids have the same handedness), the catalytic activity of proteins would have been extremely poor; it is hard to imagine how the complex structures of life could have evolved under such conditions. Similar observations apply to nucleic acids. It would appear, then, that homochirality in biomolecules must have arisen before life.

In support of this view, various autocatalytic, symmetry-breaking models, such as the Kondepudi-Nelson mechanism referred to previously, have been proposed. Yet no one has been able to pinpoint a particular set of prebiotic compounds that have all the properties required by such models. Some investigators consider this a serious difficulty; it is one of the main reasons they think chiral asymmetry must have arisen not before but after the first cell.

According to this view, the first cell developed as a singular event, and it did not possess the strongly chiral chemistry characteristic of mod-

ern life. The original "common ancestor" of all life was accidentally created with a small excess of L-amino acids or D-sugars and so incorporated only a slight chiral asymmetry. Proteins made of only one enantiomer are better catalysts, nucleic acids made of only one enantiomer are more stable and L-proteins interact more efficiently with D-nucleic acids. Therefore, in a competitive environment, evolutionary refinement of succeeding generations gradually produced life with all L-proteins and all D-nucleic acids. There is still the problem of imagining a viable life-form—the original common ancestor—made of biopolymers that contain nearly equal numbers of L- and D-enantiomers. To avoid this difficulty, some students of the subject have proposed that, by chance, the first cell already had proteins composed entirely, or nearly entirely, of L-amino acids. By any reasonable estimate, however, the probability of this happening is extremely small.

Some have proposed a third possibility—that the appearance of life was not a singular event. Symmetry breaking occurred in many places randomly, without being influenced by the chirally asymmetric weak force. In places dominated by D-amino acids, "D life" arose, and in places dominated by L-amino acids, "L life" arose. The two forms competed, and D life vanished without a trace.

Clearly, the key questions about the origin of chiral asymmetry in life remain unanswered, as do questions concerning the origin of chiral asym-

metry on a macroscopic level. Although it is now evident that the weak force, acting on the level of elementary particles, can give rise to handedness and left-right asymmetry in atoms and molecules, it is not known if these characteristics are expressed at the level of plants and animals. The chiral asymmetry in snails' shells, for example, does not appear to be related in any way to the asymmetry incorporated in their DNA or proteins; the offspring of sinistral snails can be dextral. The answers to questions about handedness in snails, human beings, cabbages and kings will have to await further revelations from developmental and evolutionary biology.

FURTHER READING

THE AMBIDEXTROUS UNIVERSE: MIRROR ASYMMETRY AND TIME-REVERSED WORLDS. Second Revised and Updated Edition. Martin Gardner. Charles Scribner's Sons, 1979.

WEAK NEUTRAL CURRENTS AND THE ORIGIN OF BIOMOLECULAR CHIRALITY. D. K. Kondepudi and G. W. Nelson in *Nature*, Vol. 314, No. 6010, pages 438-441; April 14, 1985.

PARITY VIOLATION AND THE ORIGIN OF BIOMOLECULAR CHIRALITY. Dilip Kondepudi in *Entropy, Information, and Evolution: New Perspectives on Physical and Biological Evolution*. Edited by Bruce H. Weber, David J. Depew and James D. Smith. The MIT Press, 1988.

MAPPING THE WEAK CHIRALITY OF ATOMS. R. A. Hegstrom, J. P. Chamberlain, K. Seto and R. G. Watson in *American Journal of Physics*, Vol. 56, No. 12, pages 1086-1092; December, 1988.

Stress in the Wild

Studies of free-ranging baboons in an African reserve are helping to explain why human beings can differ in their vulnerability to stress-related diseases

by Robert M. Sapolsky

The year was 1936. Hans Selye, a young physician just starting off in research at McGill University in Montreal, had a major problem. He had been injecting rats daily with a chemical extract to determine the extract's effects and had identified consistent changes in the animals: peptic ulcers, atrophy of immune-system tissues and enlargement of the adrenal glands. To his surprise, however, the rats in the control group, which had been injected with saline solution alone, showed identical changes.

Most scientists would have thrown up their hands at this paradox. Instead Selye focused on what the two groups had in common: the repeated injections. He wondered if the trio of changes he had identified was actually a generalized physiological response to unpleasantness per se.

He then tested that idea and found the same three effects regardless of whether rats were made too hot or too cold or were exposed to pathogens, toxins or loud noises. Selye borrowed a term from engineering to describe the body's nonspecific response to an insult. What the rats were undergoing, he decided, was stress. Thus, the field of stress physiology was born.

Since 1936 important details have been added to Selye's initial char-

acterization of the stress response, which is now known to involve the secretion of perhaps a dozen hormones and the inhibition of various others. Many studies have also demonstrated that chronic activation of the stress response can impair health. Moreover, some people seem to be more vulnerable to stress-related disorders than others. What accounts for the difference in susceptibility? Is it simply that some people are exposed to more stress in their daily lives, or do people actually differ in how their bodies respond to stress?

I am approaching these questions in an unusual way—by studying stress in free-ranging baboons. My ongoing research program has added strong support to a growing body of work suggesting that people's psychological and social characteristics (for example, their emotional makeup, personality and position in society) can profoundly influence their physiological response to stress.

Although chronic activation of the stress response can be harmful, few individuals could live for very long if their bodies were unable to invoke it. In fact, the stress response enables an organism to withstand immediate threats to its homeostatic balance, or physiological equilibrium. The response can be triggered by an actual insult (a physical stressor), such as extreme cold or the attack of a predator, or by the mere expectation (a psychological stressor) that an insult is about to be delivered.

In essence, the stress response prepares the body for "fight or flight." Glucose, the body's primary source of energy, is mobilized from storage sites. Blood, which transports glucose and oxygen, is diverted from organs that are not essential for physical exertion, such as the skin and intestines, and is delivered quickly to organs that are crucial—namely, the heart, the skeletal muscles and the brain. The

shift in blood flow is accomplished in part by constricting some blood vessels, dilating others and increasing the heart rate. Meanwhile cognition is sharpened (perhaps to facilitate the processing of information), and the perception of pain is blunted. And physiological activities that are not of immediate benefit are deferred; hence, growth, reproduction, inflammation and digestion—all of which are expensive, optimistic processes—are inhibited.

Chronic activation of the stress response can damage health by various means. If glucose is constantly mobilized instead of being stored, then healthy tissues atrophy, and fatigue sets in. With enough time, the cardiovascular changes promote hypertension, which in turn can damage the heart, the blood vessels and the kidneys. Moreover, when constructive processes are deferred indefinitely, the body pays a price in the form of impaired growth and tissue repair, reduced fertility and, as Selye's results suggested, diminished immune function and increased susceptibility to peptic ulcers.

As new links between stress and disease emerge, it sometimes seems miraculous that anyone can function in the modern world without being incapacitated by stress. Still, most people do just fine. The question of why this is the case has been addressed from several perspectives. Some investigators study the effects of stress on human beings directly. For example, people who have classic, hard-driving, Type A personalities have been found to be at increased risk for hypertension and heart disease. Yet the physiological events translating personality traits into diseases that take years to develop are difficult to trace in human beings, who, after all, have complex emotional lives and cannot be caged in laboratories for controlled, long-term study.

Taking a different approach, some

ROBERT M. SAPOLSKY is assistant professor of biology at Stanford University, assistant professor of neurosciences at the Stanford University School of Medicine and a research associate at the Institute of Primate Research of the National Museums of Kenya. He earned a bachelor's degree from Harvard College in 1978 and a Ph.D. in neuroendocrinology from Rockefeller University in 1984. He completed postdoctoral studies in 1987—the year he joined the Stanford faculty and was also awarded a MacArthur fellowship. When Sapolsky is not engaged in fieldwork, he studies how stress can damage brain cells and how the brain regulates the release of stress hormones.

investigators study such subjects as laboratory rats. For instance, in the 1960's Jay M. Weiss, then at Rockefeller University, showed that a sense of control or predictability can strongly influence an animal's physiology. For example, rats who receive a warning before they are exposed to an electric shock have a lesser stress response and less pathology in comparison with subjects who receive the same sequence of shocks but without warning. Yet the psychology of human beings is (hopefully) more complicated than that of rats, and so the subtlety of the psychological variables that can be studied in such animals is limited.

Captive primates are a reasonable alternative to both human and rodent subjects, but captivity, which is stressful in itself, can distort an animal's behavior and baseline measures of physiological functioning. Thus, it may compromise the

applicability of any findings to non-captive populations.

I have tried to circumvent some of the problems associated with captivity by studying olive baboons (*Papio anubis*) living freely in the Masai Mara National Reserve in Kenya. These intelligent animals are good stand-ins for human subjects in part because their primary sources of stress, like those of humans in modern society, are psychological rather than physical. Food is plentiful; the baboons spend only a few hours each day feeding. Predators are few, and infant mortality is low. With the luxury of plentiful resources and free time, the animals can devote themselves to distressing one another.

I study the males, who are quite adept at that activity. Violence itself is actually rare, but the hint of violence is ever present. Consider what can happen to a suitor who forms an association with a female in "heat," staying

close during the courtship period to prevent other males from taking his place. Often a rival male will shadow the couple for days, thereby disrupting the mating attempts of the initial suitor. The interloper may never formally provoke a fight but will inexorably maintain pressure on the courting male. It is not uncommon for these chess matches to result in surrender by the exhausted first suitor.

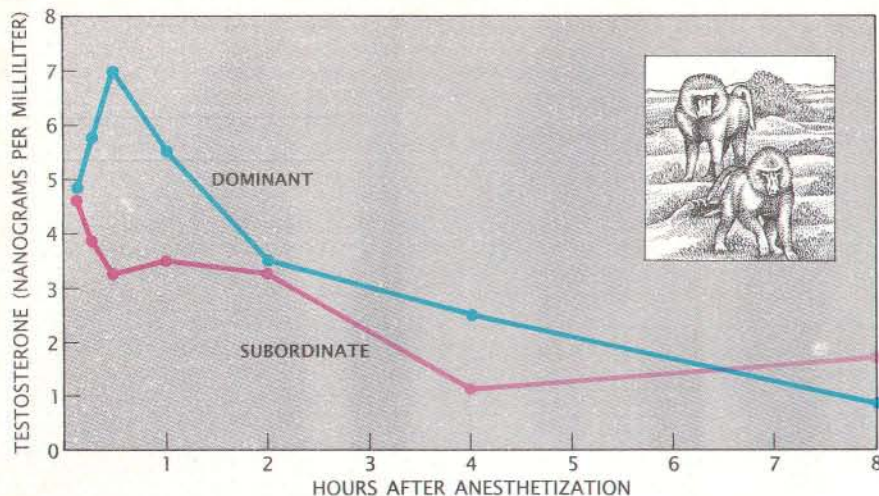
In other competitive situations, one male might form a coalition with a second male against a third. If these partnerships are stable, they can be quite successful. Long-term stability is rare, however. After spending hours establishing a coalition, a baboon may find himself abandoned in the middle of a fight or, worse, double-teamed, as his erstwhile colleague opportunistically switches sides.

Some animals are victimized more than others. The males form dominance hierarchies, and the lives of

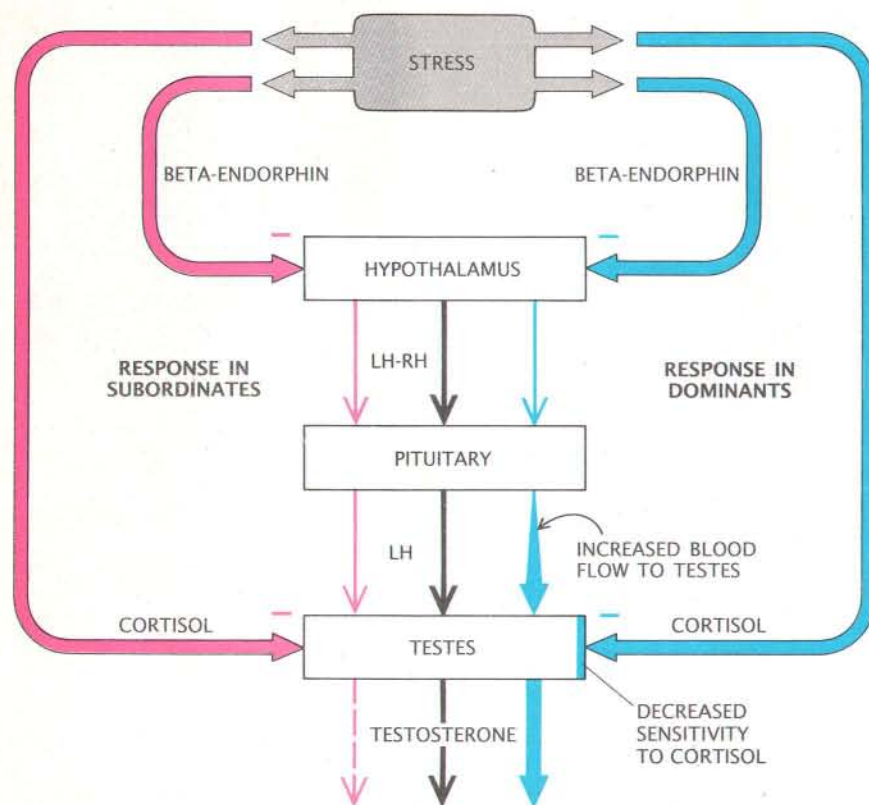


MALE OLIVE BABOON (*Papio anubis*) in the Masai Mara National Reserve in Kenya struggled to kill a gazelle for food (left) only to have his meal ended prematurely when a more dominant, or higher-ranking, male (approaching from behind) expressed interest in the bounty (right). Presumably frightened by the

interloper, the first baboon retreated quickly. Such scenes are common in the reserve; olive baboons, like human beings, are adept at stressing one another. The author has discovered that dominant males as a group generally have a different physiological response to stress than do subordinate males.



AVERAGE TESTOSTERONE LEVELS in dominant and subordinate male baboons are essentially equal when the animals are at rest but typically diverge strikingly when the animals are exposed to an identical stressor—in this case, anesthesia. The levels of the subordinate males (red) plummet immediately, whereas those of the dominant males (blue) rise sharply at first and remain elevated for approximately an hour.



CAUSES of differing testosterone levels in subordinate and dominant males during stress have been identified. When an animal rests, testosterone is released as the last step in a hormonal cascade (black arrows) beginning at the hypothalamus in the brain. The hypothalamus secretes luteinizing hormone-releasing hormone (LH-RH), causing the pituitary gland to release luteinizing hormone (LH), which in turn stimulates the testes to secrete testosterone. Stress triggers the release of beta-endorphin, an opiumlike substance, in both subordinate (red arrows) and dominant (blue arrows) males; this substance then inhibits (minus signs) the secretion of LH-RH, and thus LH, in both groups. In subordinate males testosterone levels fall because of the LH decline and because of the secretion of the hormone cortisol (hydrocortisone) during stress; cortisol tends to diminish the testes' responsiveness to LH. Testosterone levels in dominant males rise because the testes become relatively insensitive to cortisol and because the flow of blood to the testes increases; for a time, this increased flow actually increases the amount of LH that is received.

the animals who occupy the most subordinate positions are filled with a stressful lack of both control and predictability. Dominant males have easier access to food, to safe resting places and to shady spots at mid-day. They often have easier access to sexual partners and will be groomed more readily by other baboons. In contrast, subordinate males may laboriously dig tubers from the ground only to have the food nonchalantly seized by dominant males. Dominant males who lose a fight often seek a subordinate on whom to vent frustration, and they are likely to displace aggression onto the innocent bystander without warning.

Thus, the olive baboons occupy a social landscape of Machiavellian dimensions. Alliances shift unpredictably; threats range from days of harassment to sudden bursts of violence, and a baboon bent on avoiding the turmoil may still fall victim to another animal's problem.

When I began studying the olive baboons in 1978, one of my first tasks was to determine whether two baboons exposed to an identical stressor can in fact have different physiological responses. When I considered that a male's rank profoundly influences what he does in a day and how he is treated by others, I began to wonder whether rank might somehow also affect how the males respond to stress. Would the physiological response of dominant and subordinate males differ?

It turns out that the stress response does differ in the two groups. I have therefore explored the nature of these differences in some detail, along with their possible causes.

Every year I spend three months in Kenya, where I study the baboons according to a standard routine, usually with the help of a Kenyan assistant, Richard Kones. We begin by determining the males' social rank that season, which essentially involves evaluating how often the animals get what they want. For instance, we rate the animals according to whether they win most of their fights, are more often the harasser rather than the harassed, and are able to supplant another male (who might be, say, resting in a desirable spot, feeding or being groomed). I consider males in the top half of the hierarchy to be dominant and those in the bottom half, subordinate.

Once the males' social positions are known, I assess their baseline hormone levels and measure their metabolic responses to a physical stressor:

anesthesia. I anesthetize the animals by "darting" them with a syringe shot from a blowgun. Before the animals lose consciousness, they become momentarily disoriented, which seems to trigger the stress response. The anesthesia not only stresses the animals, it also makes it possible to obtain repeated blood samples over the course of the day and thus to track changes in the animals' hormone levels.

In carrying out the darting I have to adhere to many constraints. The baboons must all be injected at the same time of day, to control for rhythmic fluctuations in hormone levels. No animal can be darted if he has been injured or sick recently or if he has mated or had a major fight; such experiences will distort resting, or baseline, values of hormones. For the same reason, I have to be sure the animals have not eaten before they are anesthetized. Animals must not sense they are being stalked, or the data might be confounded by anticipatory stress. Finally, an initial blood sample (which establishes the baseline levels of the hormones to be studied) must be obtained within a few minutes after anesthesia sets in; if too much time elapses, the levels of the hormones of interest will have changed.

With this approach I have found that when the dominance hierarchy is stable (as it usually is), the workings of nearly every physiological system I have examined differ between the dominant and subordinate males. It also turns out that the physiological profile of the subordinate animals is closer to the type that is thought to predispose humans to stress-related disease.

The hormonal system that controls the secretion of testosterone (the principal reproductive hormone in males) offers a good example of how the stress response differs between dominant and subordinate olive baboons. Although the average resting levels of testosterone are essentially the same in both groups, the levels diverge markedly when the animals are stressed.

Testosterone is normally released as the final step in a cascade of hormone secretion that begins at the brain. There, the hypothalamus releases a substance known as luteinizing hormone-releasing hormone, which stimulates the pituitary gland to release luteinizing hormone. This hormone, in turn, triggers the testicular release of testosterone.

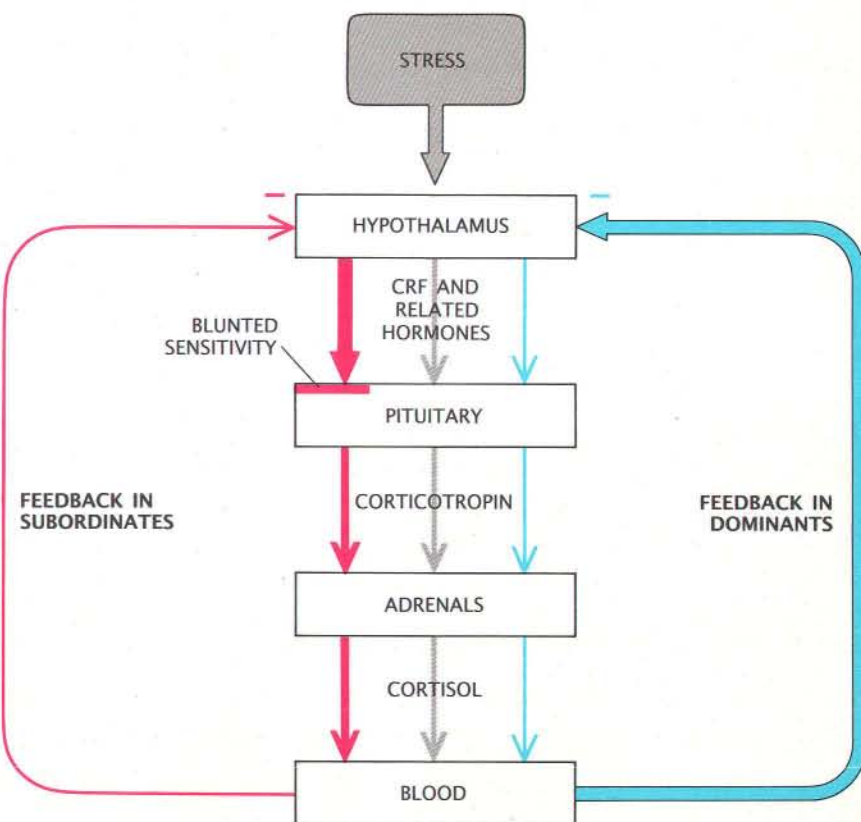
In both dominant and subordinate baboons, as in human beings and rats,

testosterone levels plummet in response to stress. Yet the similarity between dominant and subordinate males ends there. After the baboons are darted, the testosterone levels of subordinate males decline promptly, whereas those of dominant males actually rise and remain elevated for perhaps an hour before declining.

Theoretically, the rise in testosterone could give dominant males a survival and social advantage, because the hormone increases the rate at which glucose reaches the muscles. Such changes would be expected to help dominant baboons withstand a physical challenge. (Testosterone also regulates sexual behavior and aggression, but the magnitude and duration of the testosterone increase found in

dominant males during stress would not be enough to improve sexual performance or to make dominant males more aggressive than others.)

What causes testosterone levels to decline during stress, and by what mechanism are the levels elevated for a time in dominant males? I have discovered that the decline is driven in part by the stress-induced secretion of the opiumlike substance beta-endorphin, a pain suppressor that is best known for causing the so-called runner's high. Beta-endorphin, which is secreted by several organs, suppresses the hypothalamic secretion of luteinizing hormone-releasing hormone, which in turn suppresses the pituitary secretion of luteinizing hormone, leading to a decline in the



MECHANISM regulating the release of cortisol is disrupted in subordinate males, which helps explain the finding that, under normal circumstances, the mean basal cortisol levels of subordinate males are higher than those of dominant males. In both groups of animals the release of cortisol increases in response to stress (gray arrows): the hypothalamus secretes corticotropin-releasing factor (CRF) and related hormones, and these stimulate the pituitary gland to release corticotropin, which causes the adrenal glands to release cortisol into the blood. In dominant males (blue arrows) the hypothalamus receives accurate feedback from the blood, so that the brain is informed soon after a threshold level of cortisol is reached; the brain then inhibits the secretion of CRF and its relatives, leading to a decline in cortisol release. In subordinate baboons (red arrows) the feedback signal is weak, and so the brain is informed that cortisol levels are low even when they are actually high. Consequently, the hypothalamus markedly increases its secretion of CRF and related hormones. The pituitary of subordinates is somewhat insensitive to such substances, but the large amounts reaching the pituitary nonetheless trigger an increase in the secretion of corticotropin, which then leads to the chronic hypersecretion of cortisol.

amount of luteinizing hormone that reaches the testes. I determined that beta-endorphin accounts for the decline in luteinizing hormone by administering a drug to the baboons that blocks the access of the opioid to its receptors in the hypothalamus. When the activity of beta-endorphin was thus blocked, there was no stress-induced lowering of the levels of luteinizing hormone.

Another cause of the decline in testosterone levels is a decrease in the sensitivity of the testes to luteinizing hormone. This change is caused by the hormone cortisol, or hydrocortisone, which is released in quantity by the adrenal glands during stress.

The initial rise in testosterone levels after darting in dominant males cannot be explained by changes in the activity of either the brain or the pituitary gland, because the levels of luteinizing hormone released by the pituitary decline equally in high- and low-ranking males. Nor are cortisol levels involved; they are the same during stress in both groups. The explanation, then, must lie elsewhere.

I have found that a two-part mechanism seems to be responsible. In one part the testes of the dominant males somehow become less sensitive to the testosterone-inhibiting effects of cortisol. Yet, if decreased sensitivity to cortisol were the only mechanism operating, it would merely slow the decline of testosterone levels but would not lead to their elevation.

The rise itself probably results from the stress-induced release by the sympathetic nervous system of what are called catecholamines, such as adrenaline and noradrenaline, which affect blood flow. For unknown reasons, the testicular vascular system of dominant males is particularly sensitive to the dilating effects of the catecholamines, and so the testes of dominant males probably receive more blood during stress than do those of subordinate males. Hence, although the output of luteinizing hormone from the pituitary gland declines in both groups, any luteinizing hormone in the blood is probably delivered faster to the testes of dominant males. Such enhanced delivery would lead to a temporary increase in the amount of luteinizing hormone reaching the testes and so to an increase in the testicular output of testosterone.

My work has also identified rank-associated differences in the organ system responsible for increasing the release of cortisol into the blood during stress. The se-



BETRAYAL BY A COMRADE during a fight is a typical stressor for baboons. The males often form coalitions for battle but never know if a partner is reliable. In one typical

cretion of cortisol, like that of testosterone, is the final step in a cascade of hormone secretion that begins in the brain. In this case, when the animal is stressed, the hypothalamus steps up its secretion of corticotropin-releasing factor and related hormones. These hormones cause the pituitary gland to release adrenocorticotrophic hormone, also known as corticotropin. Corticotropin, in turn, stimulates the adrenal glands to release cortisol.

Cortisol is responsible for much of the double-edged quality of the stress response. In the short run it mobilizes energy, but its chronic overproduction contributes to muscle wastage, hypertension and impaired immunity and fertility. Clearly, then, cortisol should be secreted heavily in response to a truly threatening situation but should be kept in check at other times. This is precisely what occurs in dominant males. Their resting levels of cortisol are lower than those of subordinate males yet will rise faster when a major stressor does come; exactly how this speedier rise is accomplished is not understood.

I determined the cause of the higher basal cortisol levels in subordinate males by separately studying each part of the cascade that leads to the hormone's release and clearance from the blood. Working backward from the blood to the brain, I determined that the cortisol is cleared from the blood of subordinate and dominant males at the same rate. Therefore, the high cortisol levels of subordinates must stem from the excess secretion of cortisol by the adrenal glands.

This excess cortisol secretion could result from an increased sensitivity of the adrenal glands to corticotropin, excess secretion of corticotropin by the pituitary gland, or both. I found that the adrenal glands of subordinate males are not more sensitive; therefore, they must be exposed to more corticotropin.

The overproduction of corticotropin could similarly be caused by enhanced sensitivity of the pituitary gland to corticotropin-releasing factor and its relatives, excess secretion of these substances by the brain, or both. I found that the pituitary's sensitivity is actually diminished in subordinate animals. Thus, the brain probably hypersecretes corticotropin-releasing factor and its relatives, ultimately giving rise to high cortisol levels in the blood. The release of hormones by the hypothalamus cannot be measured noninvasively, but my conclusion is supported by the fact that Philip W. Gold and his colleagues at the National Institute of Mental Health came to essentially the same conclusion when they traced the causes of elevated basal cortisol levels in humans who were depressed.

Why would the brains of subordinate male baboons trigger excess cortisol release when the animals are at rest? No doubt part of the answer has to do with the animals' stressful lives, which would lead to frequent stimulation of cortisol secretion. In addition, the animals have difficulty regulating the system responsible for the secretion of cortisol.

In any chain of command, the chief needs feedback, an indication that the commands have been obeyed. In this system, the levels of cortisol—the final hormone secreted in the stress-response cascade—must be sensed by the brain. The brain should continue to evoke cortisol secretion until some threshold level of hormone has been reached, and it should inhibit secretion when the threshold is met. I wondered whether the brains of subordinates sense blood cortisol levels appropriately and found they do not.

This discovery was made by administering a synthetic version of cortisol called dexamethasone to a number of baboons. In dominant male baboons, as in most people, the brain senses the



scene two pairs of animals face off (*left*). As the fight begins, one animal abandons his partner, who is left to cope alone (*center*). Then a member of the opposing pair also withdraws, so that only two hapless combatants remain (*right*) in the end.

presence of dexamethasone and responds by curtailing the secretion first of corticotropin-releasing factor, then of corticotropin and then of cortisol. In contrast, subordinate baboons (and depressed people) are dexamethasone-resistant, that is, their brains are insensitive to the shut-off signal. As a result, cortisol production continues unchecked.

Whether subordinate males are in fact being harmed by their high basal cortisol levels remains to be seen, but certain danger signs are already evident. For instance, Glen E. Mott of the University of Texas at San Antonio and I found evidence suggesting that subordinate males may be at higher risk for atherosclerosis and thus for heart disease. In comparison with dominant male baboons, subordinates have less circulating HDL cholesterol, which is the "good" kind that helps prevent atherosclerosis. This difference was not attributable to diet, levels of activity, body weight, genetics or testosterone levels but was attributable to cortisol. We found that the higher a baboon's basal cortisol values are, the lower its levels of HDL cholesterol will be. Moreover, laboratory studies have shown that cortisol can suppress the production of HDL cholesterol.

Cortisol is known to suppress immune function during stress, and so I also compared a measure of such function in the two groups of baboons. Indeed, subordinate males have fewer circulating lymphocytes (white blood cells) than do dominant males. Although the HDL cholesterol and lymphocyte signs are ominous, the determination of whether subordinate baboons are at greater risk for heart attacks and infections can only be made by studying the same animals throughout their lives. Complicating such analyses is the fact that social rank can change over time: dominant

males wreaking havoc today may have been cringing subordinates when I first met them in 1978.

Even considering this caveat, I initially interpreted my data to suggest that the physiology of subordinate males predisposes them to stress-related disease. Rank is physiological destiny, the data seemed to say, and the other physiological systems I have studied in these males gave the same impression.

What aspect of rank might influence physiology the most? My own observations and others' studies of captive animals led me to suspect that the psychological benefits of having a high rank could be particularly important. My first hint that psychological factors might be crucial came in 1981, when the dominance hierarchy of the olive baboons became unstable. The highest-ranking, or alpha, male in my study group had passed his prime and had no heir-apparent; usually there is an obvious second-ranking animal exerting pressure on the alpha male to step aside. Instead, in this year, half a dozen young males formed a coalition to oust the alpha male. In the aftermath of the successful coup, however, the coalition disintegrated promptly. Any of these males dominated the rest of the troop's males, but among themselves, no clear hierarchy emerged. Instead months of instability ensued: coalitions formed among subgroups of dominant males and then fell apart; the amount of aggression and the number of interactions meant to test dominance increased; and ranks shifted constantly.

During this turmoil, the advantageous physiological correlates of dominance seen in other years disappeared. In contrast to males who were dominant in other study seasons, males dominant in 1981 were physiologically more like subordinates: they had elevated basal cortisol levels and sluggish secretion of cortisol in re-

sponse to stress; they also no longer had a transient rise in testosterone levels during stress. This finding suggested to me that the "better" profiles seen in dominant males in other years derived in part from the sense of control and predictability that comes with sitting atop a stable hierarchy. Although the dominant males in 1981 had the same high rank and power observed in dominant males in other years, they did not have the same sense of security.

Similar results have been found by many investigators who study captive primates, such as rhesus and squirrel monkeys. When new social groups are forming, dominant males are found to have high basal levels of both cortisol and testosterone and to be highly aggressive. Once a dominance hierarchy is stabilized, a picture emerges that resembles my sketch of the olive baboons in stable times.

The research on captive animals also indicates that the optimal hormonal profile seen in dominant males during stable times is an effect and not a cause of one's high rank. If hormonal traits accounted for dominance, the captive animals would have had different profiles even before new social groups were formed, but they did not. Thus, the beneficial physiology seen in dominant males seems to emerge from, instead of giving rise to, dominance and to arise only when dominance brings with it certain psychological advantages.

My most recent studies have altered my thinking about the influence of rank on physiology. They indicate that the advantageous physiology enjoyed by dominant males in a stable hierarchy is not a result of dominance after all. Rather, the "better" physiology found in the dominant males as a group is accounted for by a subset of animals that have certain personality traits. The traits

SCIENTIFIC AMERICAN

In Other Languages

LE SCIENZE

L. 3,500/copy L. 35,000/year L. 45,000/[abroad]

Editorial, subscription correspondence:

Le Scienze S.p.A., Via G. De Alessandri, 11

20144 Milano, Italy

Advertising correspondence:

Publietas, S.p.A., Via Cino de Duca, 5,

20122 Milano, Italy

サイエンス

Y950/copy Y10,440/year Y14,000/[abroad]

Editorial, subscription, advertising correspondence:

Nikkei Science, Inc.

No. 9-5, 1-Chome, Otemachi

Chiyoda-ku, Tokyo, Japan

INVESTIGACION Y

CIENCIA

500 Ptas/copy 5500 Ptas/year 6200 Ptas [abroad]

Editorial, subscription, advertising correspondence:

Prensa Científica S.A.,

Calabria, 235-239

08029 Barcelona, Spain

SCIENCE

27FF/copy 265FF/year 315FF/year [abroad]

Editorial, subscription, advertising correspondence:

Pour la Science S.A.R.L.,

8, rue Férou,

75006 Paris, France

Spektrum

9.80 DM/copy 99 DM/year 112.20 DM/[abroad]

Editorial, subscription correspondence:

Spektrum der Wissenschaft GmbH & Co.

Moenchhofstrasse, 15

D-6900 Heidelberg,

Federal Republic of Germany

Advertising correspondence:

Gesellschaft für Wirtschaftspublizistik

Kasernenstrasse 67

D-4000 Dueseldorf,

Federal Republic of Germany

科学

3.80RMB/copy 45.60RMB/year \$48/[abroad]

Editorial subscription correspondence:

ISTIC-Chongqing Branch, P.O. Box 2104,

Chongqing, People's Republic of China

B MIPE HAYKH

2R/copy 24R/year \$70/[abroad]

Editorial correspondence:

MIR Publishers

2, Pervy Rizhsky Pereulok

129820 Moscow U.S.S.R.

Subscription correspondence:

Victor Kamkin, Inc.

12224 Parklawn Drive,

Rockville, MD 20852, USA

TUDOMÁNY

98Ft/copy 1,176Ft/year 2,100Ft/[abroad]

Editorial correspondence:

TUDOMÁNY

H-1536 Budapest, Pf 338

Hungary

Subscription correspondence:

"KULTURA"

H-3891 Budapest, Pf. 149

Hungary

العلوم

1KD/copy 10KD/year \$40/[abroad]

Editorial, subscription, advertising correspondence:

MAJALLAT AL-OLOOM

P.O. BOX 20856 Safat,

13069 - Kuwait

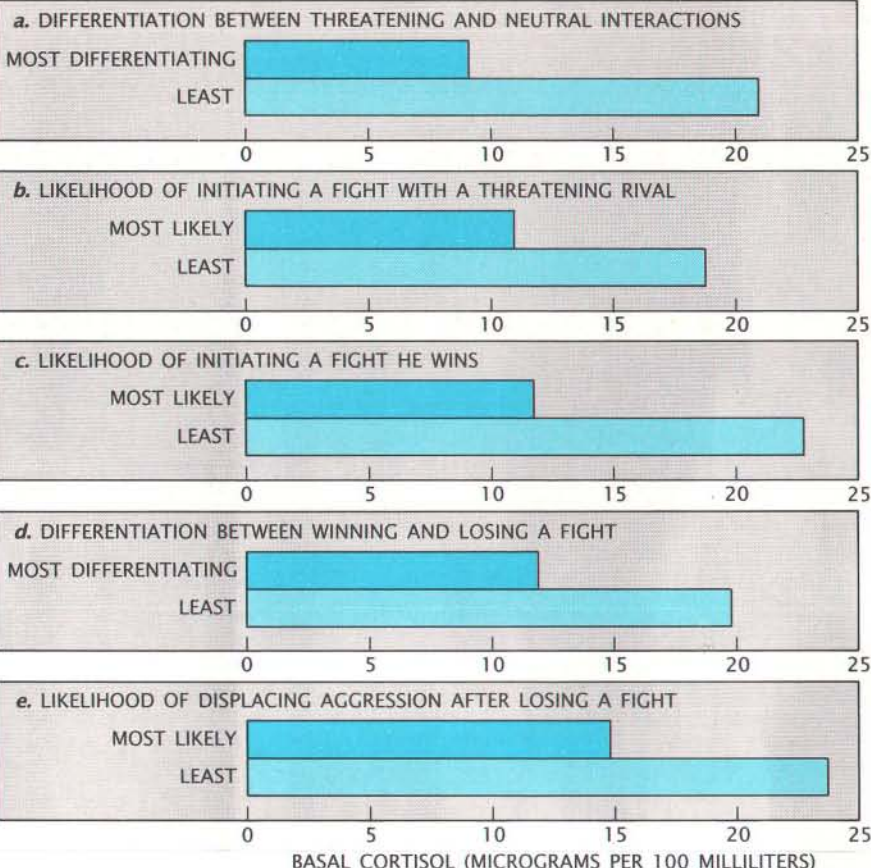
Advertising correspondence all editions:

SCIENTIFIC AMERICAN, Inc.

415 Madison Avenue

New York, NY 10017

Telephone: (212) 754-0550 Telex: 236115



DOMINANT BABOONS with certain personality traits (*dark blue*) have lower basal levels of cortisol than do other dominant males (*light blue*), which suggests that attitude is a more important mediator of physiology than is rank alone. Dominant males who can distinguish between the threatening and neutral actions of a rival have cortisol levels that are about half as high as those of other dominant males (a). Similarly, low cortisol levels are found in males who start a fight with a threatening rival instead of waiting to be attacked (b); who know which fights to pick, and so are likely to win fights they initiate (c); who distinguish between having won and lost a fight (d); or who, when they do lose, take out their frustration on subordinates (e).

apparently enable the animals to take full psychological advantage of their high rank and may, in fact, have helped the animals become dominant in the first place.

My student Justina C. Ray and I discovered the importance of personality when we analyzed the behavior of dominant males, formalizing as many distinct elements of "style" as we could imagine. (Similar studies of subordinates are now under way.) We found low basal cortisol levels—our marker of optimal physiology—only in males who have at least one of the following characteristics: they differentiate well between the neutral and threatening actions of a rival (as evinced by acting differently after each of these events); when a rival is in fact threatening, they control the situation by initiating a fight; they behave differently after winning and losing a fight; and they displace aggression onto a third party if the fight is lost.

Dominant males who lack such abilities have basal cortisol levels similar to those of subordinate males.

The general thrust of these findings is consistent with the advice routinely delivered by stress-management mavens, who say that being able to predict and control the outcome of social interactions and to find outlets for tensions can go a long way toward blunting the long-term effects of stress. The wisdom of this advice is underscored by the magnitude of the difference in basal cortisol levels between dominant males who have such traits and those who do not; the difference between these two groups is actually greater than the difference between the cortisol levels of the dominant males as a group and those of the subordinates. This finding indicates that the number of social stressors to which an individual is subjected is less important to physiology than is the emotional style with which one

perceives and copes with the stressors.

Studies of human subjects too have shown that a sense of control and outlets for distress are beneficial to physiology. For instance, in one classic study parents whose children had cancer were shown to have elevated cortisol levels. The amount of elevation varied, however, depending on the parents' coping style. Far lower cortisol levels were found in parents who had psychological defenses against anxiety, including religious faith, an ability to deny the seriousness of the child's illness, or a tendency to displace anxiety by becoming engrossed in the details of caring for the child.

It is perhaps platitudinous to conclude that attitude counts, that one must differentiate between what can and cannot be changed (and accept the latter), that one should find footholds of control and predictability in difficult circumstances. And yet my studies as well as many others have shown that stress-related physiology is remarkably sensitive to these platitudes and that the psychological filters through which external events are perceived can alter physiology at least as profoundly as the external events themselves.

For humans and animals as clever as humans, the stressors of life are predominantly socially generated ones that are both subtle and ambiguous. To the extent that so many of our stressors are the inventions of the mind, so too must be the means of coping with them.

FURTHER READING

PSYCHOENDOCRINOLOGY. Robert M. Rose in *Williams Textbook of Endocrinology*. Seventh Edition. Edited by Jean D. Wilson and Daniel W. Foster. W. B. Saunders, 1985.

SEX & FRIENDSHIP IN BABOONS. Barbara Boardman Smuts. Aldine Publishing Co., 1985.

STRESS, SOCIAL STATUS, AND REPRODUCTIVE PHYSIOLOGY IN FREE-LIVING BABOONS. Robert M. Sapolsky in *Psychobiology of Reproductive Behavior: An Evolutionary Perspective*. Edited by David Crews. Prentice Hall, 1987.

THE PSYCHONEUROENDOCRINOLOGY OF STRESS—A PSYCHOBIOLOGICAL PERSPECTIVE. Seymour Levine, Sandra G. Wiener and Christopher Coe in *Psychoendocrinology*. Edited by Seymour Levine and F. Robert Brush. Academic Press, 1989.

STYLES OF DOMINANCE AND THEIR ENDOCRINE CORRELATES AMONG WILD OLIVE BABOONS (*PAPIO ANUBIS*). R. Sapolsky and Justina C. Ray in *American Journal of Primatology*, Vol. 18, No. 1, pages 1-13; 1989.

1991 GERARD PIEL AWARD FOR SERVICE TO SCIENCE IN THE CAUSE OF MAN

Nominations are requested for the fourth Gerard Piel Award for service to Science in the Cause of Man, to be presented by the International Council of Scientific Unions (ICSU) at its 23rd General Assembly in Sofia, Bulgaria, in October of 1990. The Award, established by the Board of Directors of Scientific American, Inc., was first bestowed on Gerard Piel, creator of the magazine *Scientific American*, upon his retirement as Chairman. The Award recognizes contributions to the wise use of science for the benefit of human welfare and fulfillment. It may recognize a lifelong or an episodic contribution to this cause. The prize will consist of a sum of \$10,000 and a medal. Individuals and organizations are eligible. The Award is administered by a different scientific organization each year.

All nominations should include the following information, submitted on a typed letter: nominee's name, address, institutional affiliation and title; a brief biographical résumé, and a statement of justification for the nomination. Nominations of organizations should include information about the nature, form and work of the organization. All nominations must include the name, address, telephone number and signature of the person making the nomination.

Nominations, as well as questions about the Award, should be addressed to:

Executive Secretary
International Council of Scientific Unions
51 Blvd. de Montmorency
Paris 75016, France
Telephone: (33-1) 4525-0329
Telex: ICSU 630553 F
Telefax: (33-1) 4288-9431

Deadline for receipt of nominations is March 15, 1990.

Microplasmas

Two or more atoms—stripped of their outer electrons, trapped by electromagnetic fields and cooled to temperatures near absolute zero—array themselves in structures that behave like both liquids and solids

by John J. Bollinger and David J. Wineland

In 1973 a container whose "walls" were built from electric and magnetic fields trapped a single electron. Then in 1980 a similar device confined a single atom. The technology enabled physicists to measure the properties of electrons and atoms in unprecedented detail. The workers who initiated these experiments, Hans G. Dehmelt of the University of Washington and Wolfgang Paul of the University of Bonn, shared the 1989 Nobel prize in physics. Employing the same control over the temperature and position of atoms, we and our colleagues are investigating fundamental theories of atomic structure by trapping as many as 15,000 ions (atoms stripped of one or more of their electrons). The result is called a microplasma, by extension from the large groups of ions and electrons known as plasmas.

A microplasma is made by first applying electromagnetic fields to confine the ions to a specified region of space. A technique called laser cooling can then cool the trapped ions to temperatures of less than a hundredth of a kelvin. Because microplasmas can be built up practically one ion at a time, they provide an excellent opportunity to explore mesoscopic systems,

that is, collections of ions too small to behave like a familiar, macroscopic system and yet too complex to be identified with the behavior of a single ion. Furthermore, microplasmas can serve as models for the dense plasmas in stellar objects.

Like the atoms in liquids, the ions in some cold microplasmas can diffuse through a somewhat ordered state. In other cases, the ions can resemble the atoms in solids, diffusing very slowly through a crystal lattice. Yet the nature of microplasmas is quite different from that of conventional liquids and solids. Whereas common liquids and solids have densities of about 10^{23} atoms per cubic centimeter, microplasmas have concentrations of about 10^8 ions per cubic centimeter. Consequently, the average distance that separates ions in a microplasma is about 100,000 times greater than the distance between atoms in common liquids or solids. Furthermore, whereas internal attractive forces between the atoms hold a conventional liquid or solid together, external electric and magnetic fields hold the trapped ion microplasmas together. Indeed, the ions, which all have the same charge, actually repel each other and tend to disperse the microplasma.

called the coupling, which can be derived from particle density and temperature, describes the thermodynamic properties of a one-component plasma by providing a measure of how strongly neighboring ions interact. The coupling is defined as the Coulomb potential energy between nearest neighboring ions divided by the kinetic energy of the ions. The Coulomb potential energy depends on both the average distance between the ions (a function of density) and the charge of the ion species. The kinetic energy is simply the temperature multiplied by a physical constant known as the Boltzmann constant.

When the Coulomb potential energy is less than the kinetic energy—that is, when the coupling is less than one—the one-component plasma should have no obvious structure and should behave like a gas. But a one-component plasma whose coupling is greater than one should show some spatial order. In such strongly coupled one-component plasmas, the ions should stay away from each other because the repulsive Coulomb forces are greater than the thermal forces. At couplings of two or more, a plasma should exhibit liquid behavior. At couplings near 180, a one-component plasma should change from a liquid to a solid phase, in which the ions are arranged in a body-centered cubic crystal.

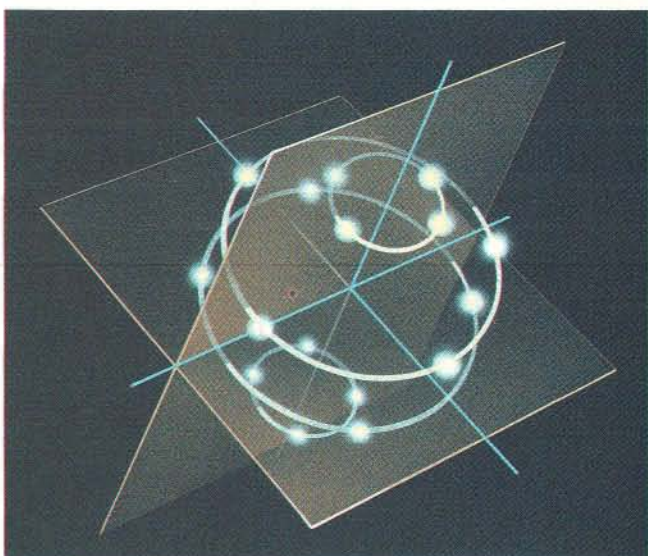
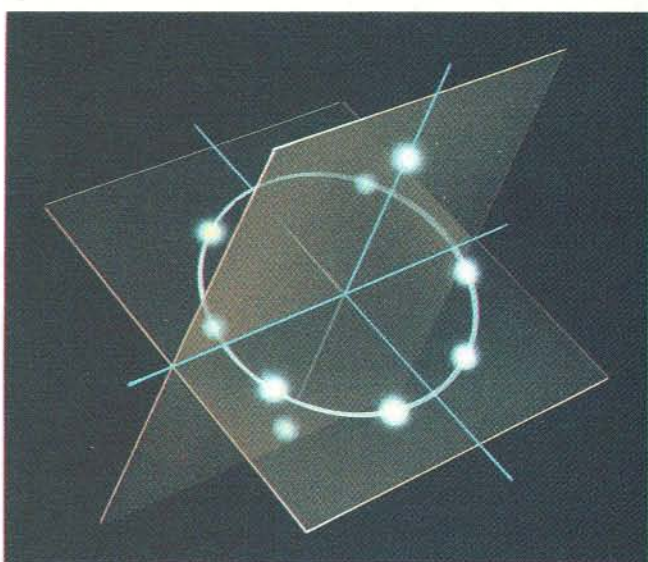
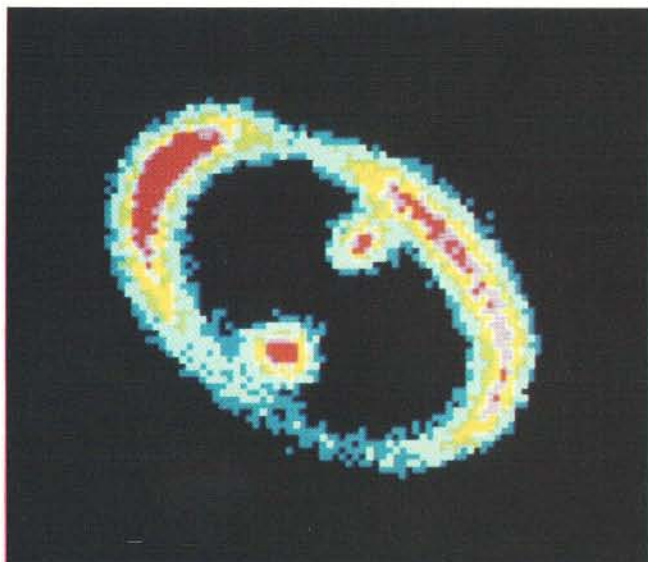
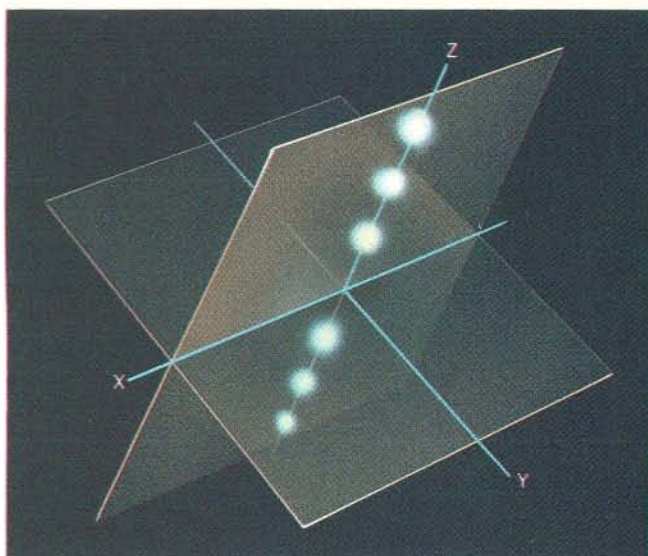
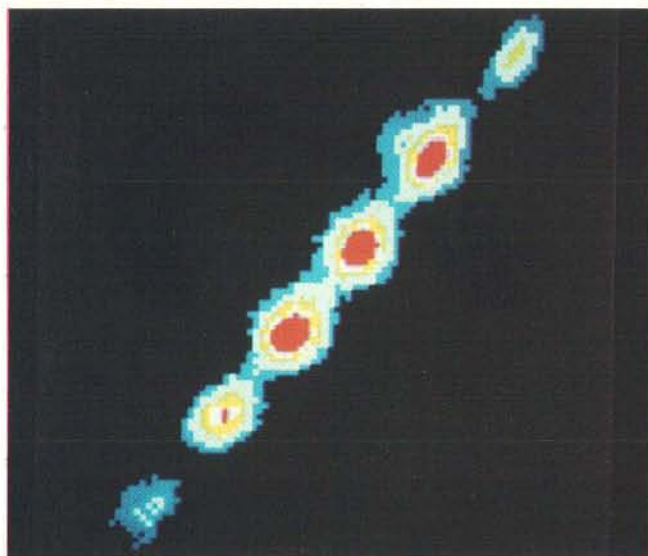
These theoretical predictions for one-component plasmas pertain to "infinite" systems, ones whose macroscopic properties do not change when a large number of ions are added or subtracted. In addition, the predictions are valid as long as the ions in the plasma behave classically, that is, as long as the effects of quantum mechanics can be neglected. Under conditions of high density and low temperature, quantum mechanics can be important, as Eugene P. Wigner first investigated in 1934.

Examples of strongly coupled one-component plasmas can be found in

JOHN J. BOLLINGER and DAVID J. WINELAND are physicists at the National Institute of Standards and Technology in Boulder, Colo. Their work on microplasmas grew out of efforts to develop techniques for high-resolution spectroscopy of stored ions. Bollinger received a B.A. in physics from Cornell University and a Ph.D. from Harvard University in 1981. He enjoys playing volleyball and hiking in the Rocky Mountains. Wineland earned a B.A. in physics from the University of California, Berkeley, and a Ph.D. from Harvard University in 1970. He was a research associate at the University of Washington and joined the Institute in 1975. An avid bicyclist, Wineland "plans to race in the Tour de France in his next life."

The first investigations of these cold plasmas began more than a decade ago. In 1977 John H. Malmberg and Thomas M. O'Neil of the University of California at San Diego suggested that a collection of electrons or ions in an electromagnetic trap would resemble a type of matter known as a one-component plasma. In such a plasma, a rigid, uniform background of charge confines mobile, identical particles of opposite charge. The specific heat, melting point and other thermodynamic properties of a one-component plasma depend greatly on the density and the temperature of the mobile particles.

A single dimensionless parameter



MICROPLASMAS composed of six, nine and 16 ions of mercury (top, middle and bottom, respectively) are held in a Paul trap. A structural diagram is shown next to each photograph. The dia-

grams are based on the predictions by Wayne Itano of the National Institute of Standards and Technology. Although the ions keep the same relative positions, they orbit around the z axis.

the universe, especially in dense stellar objects. The outer crust of a neutron star (the collapsed remnant of a large star that has exploded as a supernova) is expected to contain from 10^{26} to 10^{29} iron atoms per cubic centimeter—a density at least 1,000 times greater than anything on or within the earth. How does a strongly coupled one-component plasma form in this environment? The tremendous pressure in the star's crust breaks down the iron atoms into iron nuclei and free electrons. The iron nuclei behave classically: one positively charged nucleus simply repels its identical neighbors. On the other hand, the free electrons obey the laws of quantum mechanics, specifically the exclusion principle: every electron must occupy a different energy state. Because of the high density of electrons in a neutron star, the electrons are forced into very high energy states. The electrons are therefore unaffected by the motion of the much lower-energy iron nuclei. Hence, they form a uniform density background of negative charge. The mobile nuclei in the electron background form a one-component plasma whose coupling is estimated to range from 10 to 1,000.

To learn more about such natural plasmas, workers have attempted to generate a strongly coupled one-component plasma in the laboratory. Because the thermodynamic properties of a one-component plasma depend only on the coupling, a one-component plasma that is cool and diffuse can have the same properties as a one-component plasma that is hot and dense. Yet until recently one-component plasmas in the laboratory have not been dense enough or cool enough to become strongly coupled. In the outer crust of a neutron star, for example, the density of iron nuclei is roughly 20 orders of magnitude greater than the typical density of ions in the trap. To create a one-component plasma whose coupling matches that of a neutron star, workers must cool trapped, charged ions to a temperature that is roughly nine orders of magnitude less than that of the star. Attaining a high enough coupling demands temperatures well below one kelvin.

In addition to having a coupling equal to that of the natural system, the laboratory one-component plasma must include enough ions to reveal the behavior that is characteristic of the many ions in the natural system. Physicists have recently taken the first step and are now working on achieving the second.

Both efforts involve the technology of electromagnetic traps. The technology is roughly 30 years old: in 1959 Ralph F. Wuerker, Haywood Shelton and Robert V. Langmuir in the laboratory at Thompson Ramo-Wooldridge, Inc., in California tested an electromagnetic trap that confines charged metallic particles. This beautiful experiment demonstrates the effects caused by the strong coupling of the particles. At about the same time, workers began trapping electrons and ions [see "The Isolated Electron," by Philip Ekstrom and David Wineland; *SCIENTIFIC AMERICAN*, August, 1980].

In principle an electromagnetic trap can be nothing more than a sphere uniformly filled with negative charge. When a positively charged particle is released within the sphere, it is pulled toward the sphere's center by the uniformly distributed negative charge. Overshooting the center, the positive particle experiences a "restoring" force that is proportional to its distance from the center and eventually pulls it back toward the center again. As long as the particle is free to move through the sphere, it will oscillate about the center. If the positive particle is cooled gradually, however—that is, if its kinetic energy is decreased—it will oscillate over an ever smaller distance until it settles in the center.

Practical devices can approximate this ideal trap. Two types of electromagnetic traps in particular—the Paul trap and the Penning trap—can produce strongly coupled one-component plasmas.

The Paul trap consists of three metallic electrodes: a ring and two end caps [see *illustration on opposite page*]. The ring is wired to a generator whose voltage fluctuates sinusoidally at the so-called driving frequency. The electrodes are supported within a vacuum chamber to ensure that air molecules do not collide with the ions.

How does the Paul trap confine ions? The sinusoidal ring voltage produces a time-varying electric field between the ring and the end caps. The direction of the field alternates: half the time the ion is pushed toward the ring and pulled away from the end caps, and half the time the ion is pulled away from the ring and pushed toward the end caps. The ion wiggles sinusoidally at the driving frequency.

Because the driving force varies sinusoidally over time, one might expect the net force on the ion to be zero. In other words, one might think that the ion would move back and

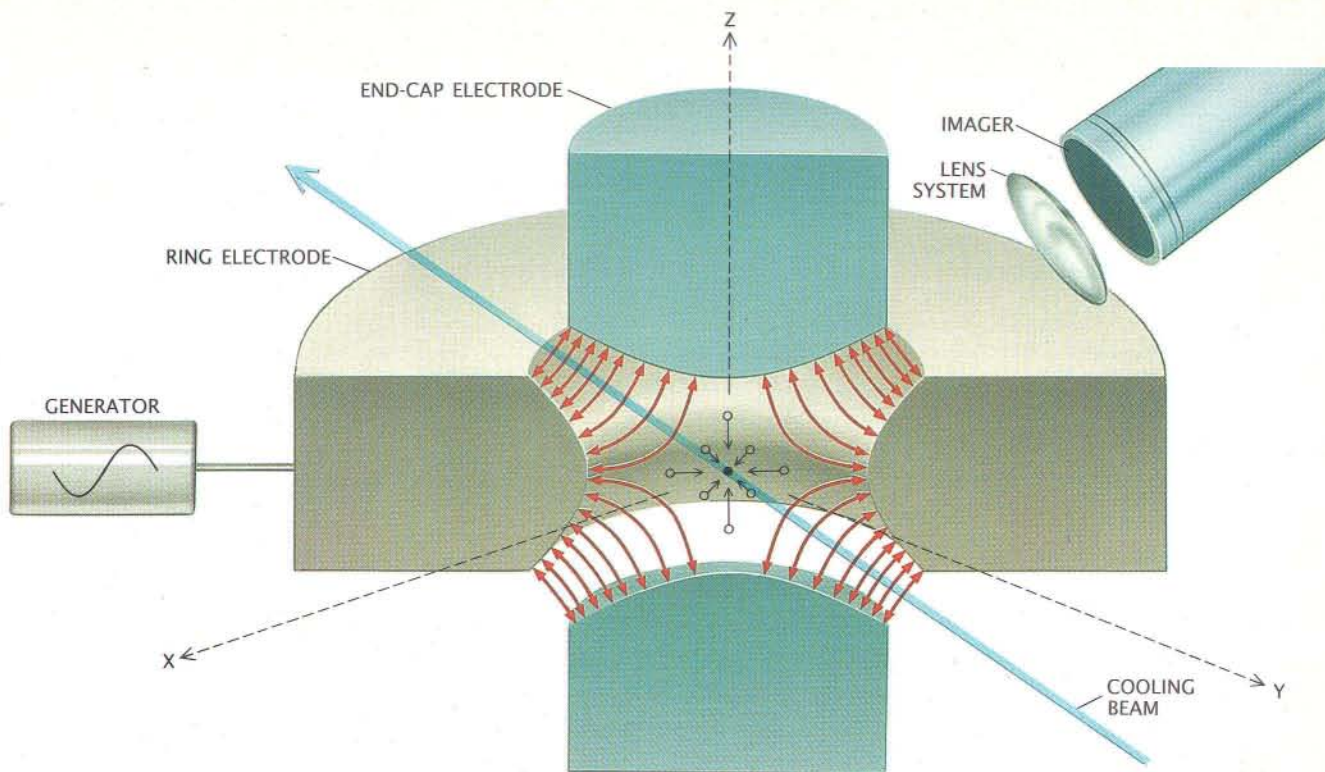
forth under the influence of the oscillating electric field but that the center of its oscillation would remain fixed over time. What, then, drives the ion toward the center of the trap?

The answer stems from the fact that because of the shape of the electrodes, the electric field is weaker at the center of the trap than it is near the electrodes. By considering a special case, one can gain a basic idea of how the spatial variation in the field affects an ion. Suppose that for a time the forced sinusoidal motion of an ion is centered closer to the top end cap than to the bottom end cap. When the ion is at the top of its oscillation, it encounters a strong force toward the center of the trap. When it is at the bottom of its oscillation, the force is toward the top end cap but is somewhat weaker. The ion therefore experiences a net force toward the center of the trap—the so-called ponderomotive force.

In the same way, the ion will experience a ponderomotive force that tends to drive it toward the center of the trap if it oscillates below the center or to either side. In the vertical direction the ponderomotive force is called the axial force; in the lateral direction it is called the radial force.

At all times, then, an ion in the trap experiences a driving force and a ponderomotive force. It turns out that the axial ponderomotive force is greater than the radial ponderomotive force. An ion therefore oscillates with three characteristic frequencies: driving, radial and axial. The trap is usually designed so that the radial and axial ponderomotive frequencies are about 10 times lower than the driving frequency. Hence, the motion caused by the driving force is a small, fast wiggle superposed on a large, slow oscillation about the center of the trap that is caused by the ponderomotive force.

If one disregards the smaller, faster oscillation caused by the driving force, an ion in the Paul trap moves in the same way it would inside a negatively charged sphere. Because the ponderomotive force differs in the axial direction, however, it is more appropriate to think of the Paul trap as a spheroid that is oblate (pancake-shaped) or prolate (cigar-shaped). To control the shape of the spheroid and the combined forces on the ions, one can apply an additional constant voltage between the ring and the end caps. When the overall radial force is greater than the axial force, the trap acts as a prolate spheroid. Conversely, when the radial force is less than the axial force, then the trap acts as an oblate spheroid.



PAUL TRAP creates a time-varying electric field (red lines) between the electrodes. As a generator applies a changing voltage to the ring, the electric field changes strength and direc-

tion, but its shape stays constant. The resulting ponderomotive forces (black arrows) confine charged particles. A laser beam directed at the trap's center cools and probes the ions.

oid. Thus, the Paul trap can both confine the ions in the center and orient the collection in the axial or radial direction.

In 1987, working at the National Institute of Standards and Technology with James C. Bergquist, Wayne M. Itano and Charles H. Manney, we used a Paul trap to observe strongly coupled microplasmas of mercury ions. At the same time, groups led by Herbert Walther and Frank Diedrich of the Max Planck Institute for Quantum Optics in Garching, by Peter E. Toschek of the University of Hamburg and by Richard G. Brewer of the IBM Almaden Research Center in San Jose, Calif., were conducting similar experiments on various ion species.

At the start of our experiment, a small amount of mercury vapor was allowed to leak into the vacuum system containing the Paul trap. As the mercury atoms passed through the trap, they were bombarded with a beam of electrons. The electrons in the beam had just enough energy to knock a single electron out of any mercury atom they struck, thereby ionizing the atom.

To cool the ions, we relied on the technique of laser cooling [see "Cool-

ing and Trapping Atoms," by William D. Phillips and Harold J. Metcalf; *SCIENTIFIC AMERICAN*, March, 1987]. We generated a laser beam at a frequency slightly below a frequency that the ions could easily absorb. The ions traveling toward the laser source, however, "saw" the frequency as being slightly increased because of the Doppler effect. These ions absorbed the light strongly and slowed down. The ions traveling away from the source encountered the light at a lowered frequency; as a result, they absorbed the light weakly and did not speed up much. Overall the average motion of the ions was reduced: the ions were cooled.

To observe the individual mercury ions and their spatial structures, we illuminated them with ultraviolet light, which mercury ions scatter strongly. An ultraviolet video camera recorded up to 100,000 photons of ultraviolet light per second and thereby generated a motion picture of the trapped ions. The camera could resolve details as small as one micron (one millionth of a meter).

At first we worked with only one mercury ion in the trap and cooled it to millikelvin temperatures. The ponderomotive force confined the ion to the center of the trap. When we al-

lowed two ions into the trap, we discovered two possible configurations. If the radial ponderomotive force was stronger than the axial force, the two ions lined up in the axial direction, so that they were equidistant from the trap center. Conversely, if the radial ponderomotive force was weaker than the axial force, the ions lined up in the radial plane equidistant from the trap center. As we added more ions to the trap, our intuition about the locations of the ions became questionable. But, with the assistance of a computer to keep track of the many forces and ions, Itano simulated various conditions in the trap and accurately predicted the resulting configurations of ions.

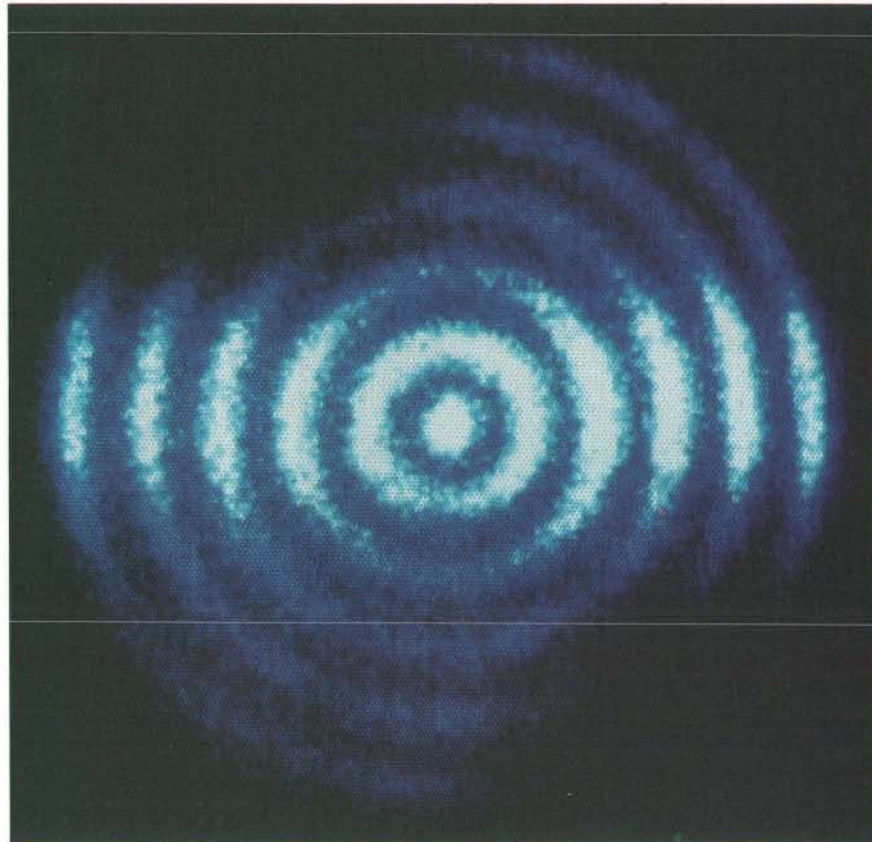
By confining ions, we confirmed that a Paul trap could support a one-component plasma. Although the ions were strongly coupled, as the observed spatial structures demonstrated, we needed to calculate the coupling from measurements of density and temperature. We could easily determine the density of the ions from images, but we had to measure temperature indirectly. We observed (as have many other groups in other experiments) that the motion of the ions modified the absorption

spectrum associated with the ions. (The absorption spectrum of an atom or molecule reveals the frequencies of radiation absorbed most strongly by the atom or molecule.)

A collection of absolutely stationary ions would have a very sharp absorption spectrum, indicating absorption only at well-defined frequencies. On the other hand, if the ions moved around to some degree, the absorption spectrum would be blurred. The blurring results from the motion of the ions toward or away from the radiation source. From the perspective of the ions, however, it is the source that is moving toward or away from them, and so the frequency of the light is shifted by the Doppler effect. Thus, ions moving toward the source will be able to absorb radiation of slightly lower frequencies more effectively than motionless ions can, and ions moving away from the source will be able to absorb radiation of slightly higher frequencies. The combination of many ions moving in many directions has the effect of "smearing out" the spectrum, and the amount of smearing discloses the temperature of the ions. This technique proved that the temperature of the ions in our trap approached 10 millikelvins. The couplings, then, were as large as 500.

These measurements of the couplings, which indicated that the Paul trap could support a strongly coupled one-component plasma, were done with fewer than about 25 ions in the trap. We found it difficult to create similar solid states with more ions. The difficulty arose from the effects of the oscillation induced in the ions by the driving force. As an ion oscillates at the driving frequency, the repulsion between the ion and its neighbors enables it to influence the oscillations of the other ions. This additional perturbation can cause a plasma's structure to heat to a breaking point under certain operating conditions of the trap. The effect is known as radio-frequency heating, because the driving frequency for atomic ions is about the same as the frequency of radio waves. Radio-frequency heating was first observed in the experiments of Wuerker, Shelton and Langmuir.

More recently Reinhold Blümel and co-workers at the Max Planck Institute and John A. Hoffnagle and colleagues at the IBM Almaden Center have studied radio-frequency heating of atomic ions in great detail. These studies show that a small change in the system parameters of the Paul trap can cause a sudden transition between cold, crystalline states and hot, gas-



eous states. For example, the rate at which the ions are cooled is very sensitive to the frequency of the laser light employed for cooling. If the laser is tuned far below the optimal cooling frequency, radio-frequency heating transforms the ion plasma into a hot, disordered state. If the laser frequency is increased, the rate of laser cooling increases, until at a critical frequency the laser cools the ion plasma quickly enough, so that some order starts to appear. At that point the radio-frequency heating diminishes drastically, and the ion plasma suddenly freezes into an ordered state. This frozen state is quite stable and will persist even if the laser frequency is again decreased somewhat. The irregular nature of radio-frequency heating makes it difficult to study the liquid-to-solid phase transitions predicted for one-component plasmas.

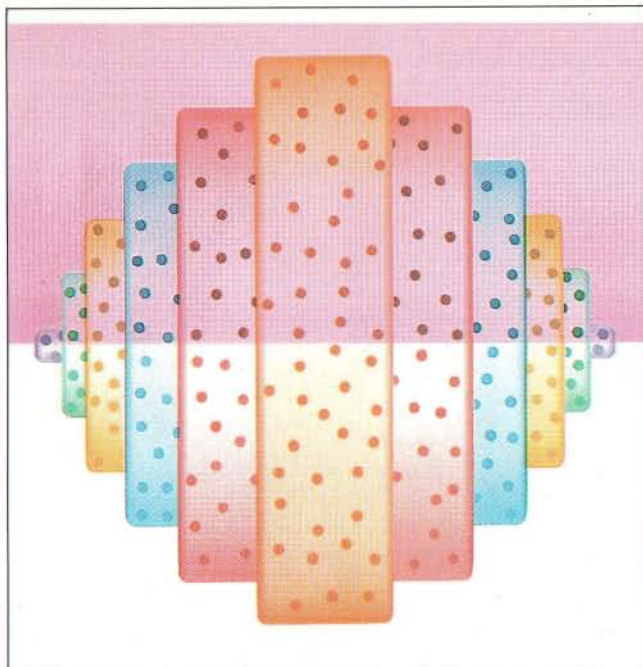
The problem of radio-frequency heating grows as the number of ions in the trap increases. When many ions are in the trap, some are pushed toward the electrodes, where the driving force is stronger. Those ions then oscillate at the driving frequency with a large amplitude and thereby increase the effects of radio-frequency heating. These considerations have limited the number of ions that can be cooled at one time in a Paul trap to about 200. If

the difficulties associated with radio-frequency heating can be overcome, the Paul trap should allow workers to study the liquid-to-solid phase transition as well as other properties of the "infinite" one-component plasma.

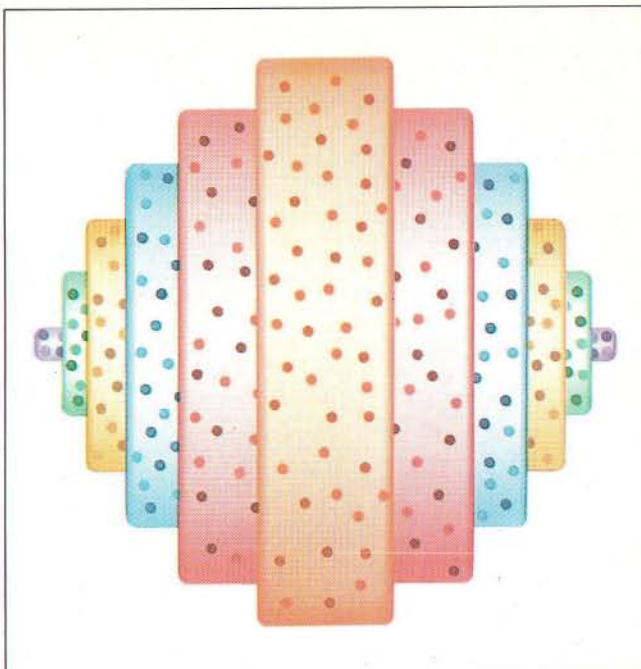
At present the Penning trap provides a more hospitable environment for experimenting with large strongly coupled one-component plasmas than the Paul trap does. Unlike the time-varying electric fields of the Paul trap, the electric and magnetic fields that confine charged particles in the Penning trap are static. In 1988, with Sarah L. Gilbert, we constructed a Penning trap to confine beryllium ions. Malmberg and colleagues have also employed a Penning-type trap to confine strongly coupled plasmas of electrons.

Our trap consisted of four cylindrical electrodes arranged end to end along a common axis [see illustration on page 120]. A positive voltage was applied to the two outer cylindrical electrodes. This voltage generated an electric field between each inner electrode and the adjacent outer electrode. These fields trapped the ions in the axial direction, near a plane between the inner electrodes.

A powerful magnet placed around the electrodes created a uniform mag-



SHELL STRUCTURE of a microplasma consisting of about 1,000 beryllium ions is revealed in a photograph (left) made by shining a laser beam through the microplasma and recording the resulting fluorescence. The illustrations depict the complete



shell structure as well as the diffusion of the beryllium ions from one moment (center) to the next (right). The microplasma behaves like both a solid and a liquid: the ions (colored dots) travel around within the shells but not between the shells.

netic field directed along the axis of the cylinders. The magnetic field prevented the ions from leaving the trap in the radial direction. The radial force of the electric field near the center of the trap is directed away from the trap center. This force combines with the axial magnetic field to cause the ions to orbit about the trap axis. As the orbiting ions pass through the magnetic field, they experience a Lorentz force that is directed radially inward. The Lorentz force is what confines the ions radially.

Except for this uniform rotation of the microplasma, the confining forces of the Penning trap are equivalent to the confining forces in the uniformly charged spheroid. Hence, even though ions in a Penning trap rotate, they behave like ions in a one-component plasma—in particular, they have the same thermodynamic properties.

To begin the experiment, we produced beryllium ions by a method similar to that described for making mercury ions in the Paul trap. The ions were cooled by two intersecting laser beams. A third laser beam, called the probe, was employed to measure the temperature of the ions. The light scattered by the ions was collected to make an absorption spectrum. As in the mercury-ion experiment, the temperature of the ions could be deduced

from the blurring of certain features of the spectrum. This technique revealed that the ions had been cooled to below 10 millikelvins.

The rotation frequency of the beryllium microplasma could also be deduced from the spectrum. Beryllium ions circulated inside the trap at a rate of from 20,000 to 200,000 rotations per second. Because the rotation frequency is directly related to the radial electric fields, which are in turn related to the ion density, we were able to calculate that the ion density ranged from 50 to 300 million ions per cubic centimeter. From ultraviolet images of the plasmas we could determine the volume occupied by the trapped ions and therefore the number of trapped ions. The temperature and density measurements yielded couplings as large as 200 to 400 for less than about 15,000 ions in the trap.

We expect a phase transition from a liquidlike to a solidlike state to occur at a coupling of 180 for a one-component plasma with an infinite number of ions. Our measured values for the coupling indicate that the trapped ions should form a crystalline ion solid, if the trap contains enough ions. Should a system of 15,000 trapped ions, though, behave like an infinite system?

Recently some computer simulations have elucidated this question. The late Aneesur Rahman of the University of Minnesota at Minneapolis, John P. Schiffer of the Argonne National Laboratory, Hiroo Totsuji of Okayama University and Daniel H. E. Dubin and O'Neil of the University of California at San Diego have created simulations of a trapped plasma containing as many as several thousand ions. The simulations reveal several remarkable features. When the couplings exceed one, the ions are concentrated in concentric shells, which are spaced evenly. For couplings around 10, the shells are in a liquid state characterized by short-range order and diffusion in all directions. As the coupling increases, the shells become more clearly defined; the ions diffuse quickly within the shells and slowly between the shells. For high couplings (above 200), the diffusion of the ions within a shell slows down and the ions form a solidlike state. Instead of showing a sharp phase transition, the plasma evolves gradually from a liquidlike to a solidlike state.

Experiments have confirmed these predictions. Even though the plasma of beryllium ions rotates around the axis of the Penning trap, the shell structure is preserved in the radial direction. The scattered light from

a laser beam cutting across the plasma shows alternating bright and dark bands corresponding to the shells. We looked for shell structure in plasmas that contained as few as 20 ions and as many as 15,000 ions. In a plasma of 20 ions, a single shell was clearly observed. In a plasma of 15,000 ions, we

could distinguish 16 shells. So far we have not detected any distinct structure within a shell because of the rotation of the plasma in the trap.

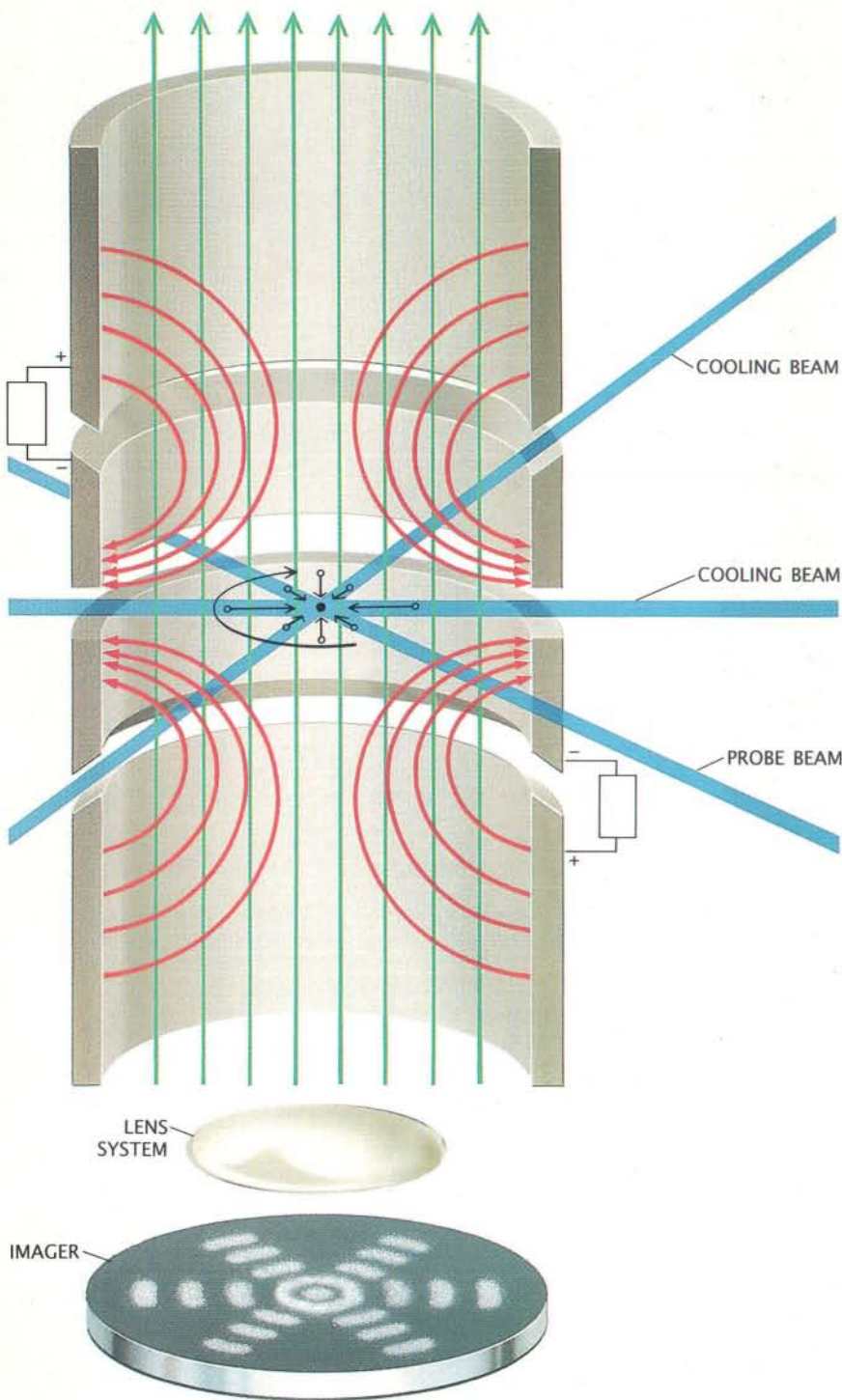
We were able to test predictions that for couplings around 100, the plasma will act like a liquid within a shell but like a solid between the shells. In par-

ticular, the ions should diffuse faster within shells than between shells. To demonstrate the effect, we optically "tagged" the ions by tuning the probe laser to a specific frequency. The probe laser suppresses the emission of light from the ions it strikes by placing them in a "dark" energy state in which they do not scatter light from the cooling laser beams.

First we darkened an outer shell of the plasma and measured the time required for the dark ions in the outer shell to move to the inner shells. Then we darkened part of the plasma across several shells and measured the time required for dark ions in one part of a shell to move to other parts of the same shell. These measurements verified that for moderate couplings the diffusion of ions between shells is more than 10 times slower than the diffusion of ions within a shell.

Many questions about microplasmas are still unresolved. At what point will the behavior that is characteristic of an infinite system start to appear? How many ions are required for the system to exhibit a sharp phase transition? At what stage will the solid state become a body-centered cubic lattice rather than a collection of shells?

At present these questions are difficult to answer even in theory. Dubin predicts, however, that perhaps as many as 50 to 60 shells may be required before a body-centered cubic lattice would become an energetically favorable configuration. That would require about a million ions, more than 50 times the number of ions in the largest, strongly coupled microplasmas created so far. Current technology should be able to confine cold plasmas of this size. If couplings of 200 or more can be maintained for a plasma of a million ions, we may be able to visit the surface of a neutron star in a laboratory on the earth.



PENNING TRAP generates electric fields (red lines) and magnetic fields (green lines) to produce forces (black arrows) that confine charged particles. The electric and magnetic fields also cause the microplasma to rotate. Two laser beams cool the particles; a third serves as a probe for various experiments. The magnetic field arises from an electric current flowing through a solenoid (not shown) that encircles the trap.

FURTHER READING

NON-NEUTRAL PLASMA PHYSICS. Edited by C. W. Roberson and C. F. Driscoll. AIP Conference Proceedings 175, American Institute of Physics, March, 1988.

PROCEEDINGS OF THE ELEVENTH INTERNATIONAL CONFERENCE ON ATOMIC PHYSICS, JULY 4-8, 1988, in *Atomic Physics 11*. Edited by S. Haroche, J. C. Gay and G. Grynberg. World Scientific Publishing Co., 1989.

PROCEEDINGS OF THE YAMADA CONFERENCE ON STRONGLY COUPLED PLASMA PHYSICS, AUGUST 29-SEPTEMBER 2, 1989. Edited by Setsuo Ichimaru. Elsevier Science Publishers, in press.

How to get to Italy once a month.

If you want your advertising to reach industry and government leaders in Italy, reach for SCIENTIFIC AMERICAN's Italian edition, **LE SCIENZE**.

Today, Italy has mastered the finest points of high technology across the full spectrum of industries. The people responsible for Italy's achievements on the technology front are reading **LE SCIENZE**.

Tony Severn will give you details:

LE SCIENZE S.p.A.
Piazza della Repubblica, 8
20121 Milano, Italy
Telephone (011) 39-2-655-4335
Telefax (011) 39-2-655-2908

Circulation 72,000



The Cosmic Background Explorer

NASA's cosmological satellite will observe a radiative relic of the big bang. The resulting wealth of data will be scoured for clues to the evolution of structure in the universe

by Samuel Gulkis, Philip M. Lubin, Stephan S. Meyer and Robert F. Silverberg

Late last year the National Aeronautics and Space Administration launched its first satellite dedicated to the study of phenomena related to the origins of the universe. The satellite, called the *Cosmic Background Explorer* (COBE), carries three complementary detectors that will make fundamental measurements of the celestial radiation. Part of that radiation is believed to have originated in processes that occurred at the very dawn of the universe. By measuring the remnant radiation at wavelengths from one micrometer to one centimeter across the entire sky, scientists hope to be able to solve many mysteries regarding the origin and evolution of the early universe.

The COBE data will be analyzed for clues to questions of the most fundamental nature: What were the conditions when the remnant radiation was emitted? How did the structures we see in the sky today develop? What was the cosmos like when the first luminous bodies formed? Can we see the diffuse radiation from a possible first generation of stars? Was there

an era when masses of intergalactic dust absorbed much of the early starlight?

The cosmic radiation, created early in the evolution of the universe under drastically different conditions than those prevailing today, probably has several parts. Each part would have originated at a different stage in the evolution of the universe, and each would have been the result of different processes. The most well-established component is the cosmic microwave background (CMB) radiation, discovered nearly 26 years ago; the measurement of its properties has been an active area of research ever since. Another component is expected to be found in the infrared region of the spectrum. This cosmic infrared background (CIB), whose existence has not yet been confirmed, is the predicted consequence of the formation of the first objects from primordial material.

Unfortunately, these radiative relics of the early universe are weak and veiled by local astrophysical and terrestrial sources of radiation. The wavelengths of the various cosmic components may also overlap, thereby making the understanding of the diffuse celestial radiation a challenge. Nevertheless, the COBE instruments, with their full-sky coverage, high sensitivity to a wide range of wavelengths and freedom from interference from the earth's atmosphere, will constitute for astrophysicists an observatory of unprecedented sensitivity and scope. The interesting cosmic signals will then be separated from one another and from noncosmic radiation sources by a comprehensive analysis of the data.

The COBE mission has been profoundly shaped by the current understanding of the universe. The discovery of the CMB has led to wide acceptance of the hot big-bang

theory, a remarkable synthesis of various observations that includes the expansion of the universe and its hydrogen-to-helium ratio, as well as the CMB itself. This theory asserts that the universe started from a primeval fireball, an extremely dense and hot state with a tiny volume, that has since expanded to its present scale. As it expanded, the matter and radiation cooled from temperatures so high that the behavior of the primordial "stuff" that existed in the first instant of the universe is beyond the predictive power of today's physics.

The cooling initiated events that led to the present universe. Neutrons and protons formed from their quark constituents. A few minutes later, nuclei of helium, deuterium and lithium coalesced from the protons and neutrons. Approximately 300,000 years after the big bang, as the universe cooled further, the nuclei combined with free electrons to form electrically neutral atoms. The initial formation of neutral atoms, referred to by cosmologists as the decoupling era, was of crucial importance to the development of stars and galaxies: it allowed the matter and radiation to evolve independently for the first time.

The radiation continued to cool as the universe expanded, so that today we observe it as the cosmic microwave background, whose properties closely resemble those of an ideal thermal source, called a blackbody, at a temperature of 2.7 kelvins. Even though it is now faint and cool, the CMB remains the dominant form of radiant energy in the universe today. Because the matter essentially stopped interacting with the radiation at the decoupling era, the present-day radiation gives us a "snapshot" of what the conditions were like when the universe was only about 300,000 years old.

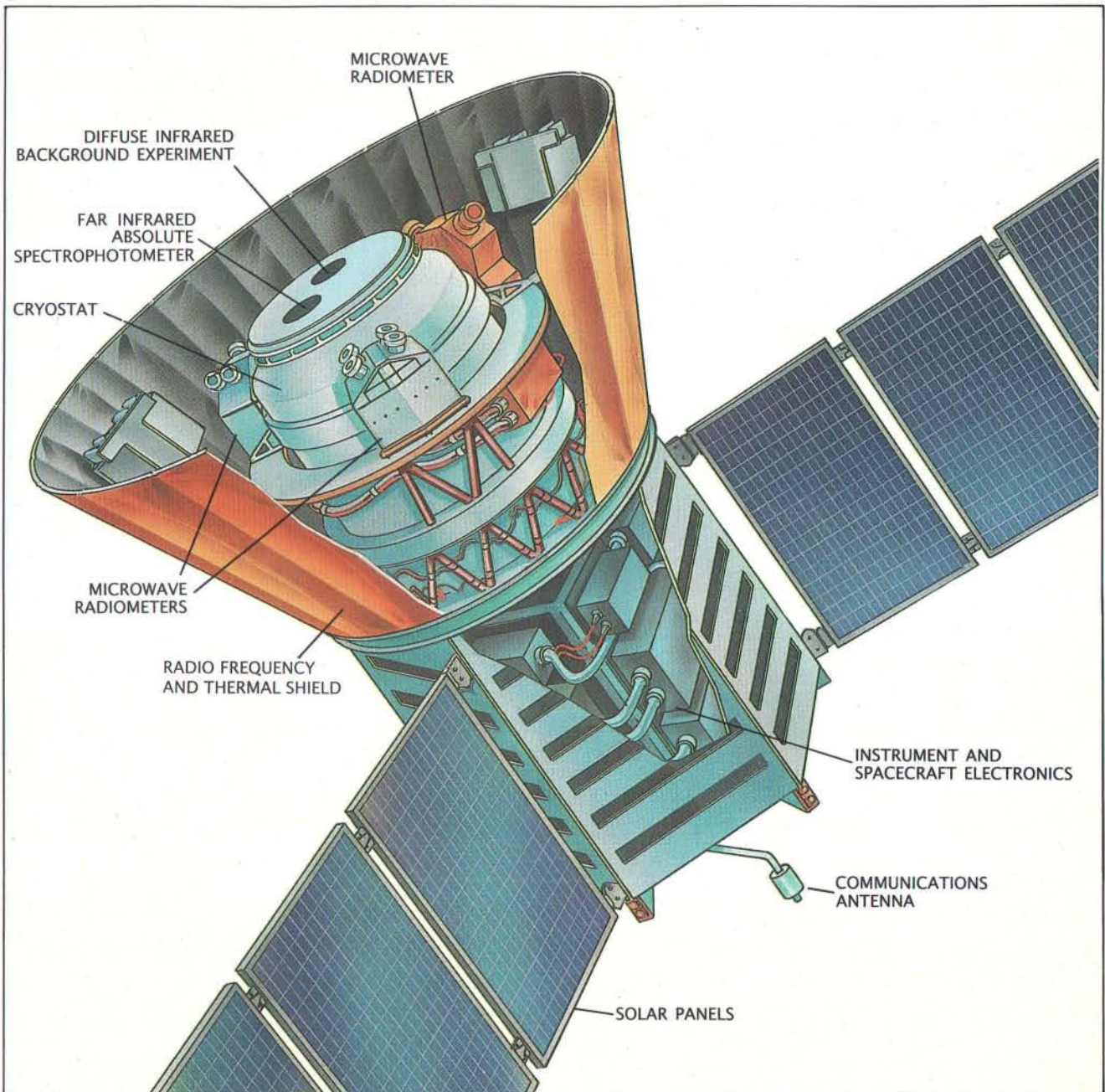
After the era of decoupling, matter evolved unhindered by radiation pressure and, under the influence of gravi-

SAMUEL GULKIS, PHILIP M. LUBIN, STEPHAN S. MEYER and ROBERT F. SILVERBERG are all co-investigators on the COBE mission. Gulkis, a senior research scientist at the Jet Propulsion Laboratory of the California Institute of Technology, has been on the COBE team since its inception in 1974. He holds a Ph.D. in physics from the University of Florida. Lubin, professor of physics at the University of California, Santa Barbara, received a Ph.D. in physics from the University of California, Berkeley. Meyer, assistant professor of physics at the Massachusetts Institute of Technology, became interested in experimental cosmology while studying for his Ph.D. at Princeton University. Silverberg is an astrophysicist at NASA's Goddard Space Flight Center. He holds a Ph.D. in physics from the University of Maryland at College Park.

ty, collapsed into the celestial objects that we now see. Because significant amounts of elements heavier than helium have been observed in the oldest known stars, the material most likely was produced in an even earlier generation of stars that also formed during this collapse period. The en-

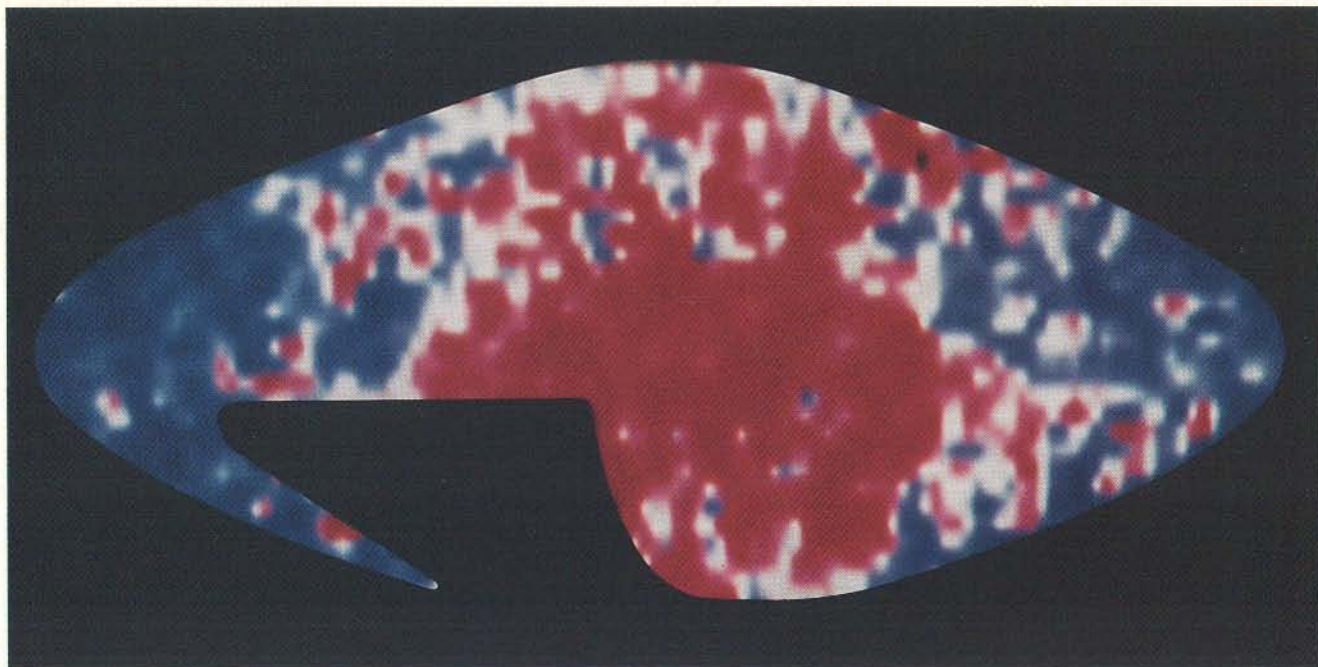
ergy from the gravitational collapse and the production of the first heavy elements must have produced a considerable amount of radiation that should now appear in the infrared. Hence, theory says, a cosmic infrared background must exist, although it has not yet been detected.

Although the scenario we have outlined is in accordance with existing measurements, it is hardly complete. Cosmologists do not know the conditions that prevailed between the initial moments of the universe and the formation of the most distant objects we now observe, a span of approximately



COSMIC BACKGROUND EXPLORER (COBE) scans the sky from an almost polar orbit near the earth's day-night terminator. A shield protects its three instruments from direct solar and terrestrial radiation. A year's supply of liquid helium, stored in the cryostat (a vacuum-insulated bottle), chills the far infrared absolute spectrophotometer (FIRAS) and the diffuse infrared background experiment (DIRBE) to improve sensitivity and reduce systematic errors. The FIRAS examines 1,000 celestial regions for evidence of energetic processes in the early uni-

verse. The DIRBE measures the absolute brightness of the sky for the vestige of the earliest starlight, a predicted but unverified phenomenon that might help explain the evolution of cosmic structure. A third set of observations is conducted by a set of three differential microwave radiometers (DMR's) deployed outside the cryostat. Each seeks to discover tiny spatial variations, or anisotropies, in the microwave intensity that might indicate whether matter was distributed unevenly at the time the cosmic microwave background radiation originated.



VAST DIPOLE in the sky appears in this nearly complete map of data collected by a balloon-borne experiment conducted by Lubin and his colleagues. The cooler regions (*blue*) have been redshifted by .1 percent, and the hotter region (*red*) has been

blueshifted to the same degree, indicating that the earth is approaching Virgo at 300 kilometers per second. That implies a galactic velocity of 600 kilometers per second with respect to the cosmic background—about .2 percent of the speed of light.

one billion years. Still, two fundamental properties of the CMB provide clues to the conditions in the early universe: its spectrum and the way it varies in intensity with respect to direction in the sky—its angular variation. Because experiments during the past 25 years or so have shown that the CMB spectrum is close to that of a blackbody source, it is clear that the universe was nearly in a state of thermal equilibrium at the period of decoupling. This observation constrains theories of the early universe. Small departures from a blackbody spectrum would imply the existence of major energetic processes before the decoupling era that would have disrupted the thermal equilibrium. Such variations might also be the consequence of matter altering the radiation in the epochs that followed the period of decoupling.

The second notable characteristic of the cosmic microwave background, the variation of its intensity from one direction in the sky to another, provides clues that may help answer other questions. Such variations, called anisotropies, could be produced by a "lumpy" distribution of matter and energy at the time of decoupling or by relative motions between the different parts of the universe. Because large-scale structures such as galaxies

would have taken a long time to coalesce if the initial distribution of matter had been uniform, theorists predict that the CMB will be found to be slightly anisotropic. Searches for CMB anisotropy that could be attributed to the "seeds" of present-day structures have been extensive and acutely sensitive, and yet results have been negative so far: on angular scales of several arc minutes, we know that the CMB is smooth to less than 20 parts per million. If one assumed that gravity is the only important force driving the evolution of structure on a large scale, it would be difficult to reconcile this smoothness in the CMB with the lumpiness of the visible matter.

One angular variation in the cosmic microwave background, believed to be associated with the velocity of the earth with respect to the radiation, has been detected. Along one direction in the sky, the CMB appears warmest; in the opposite direction it is coolest. The contrast is minute—only .1 percent warmer in the direction of the earth's motion and .1 percent cooler in the opposite direction. This angular variation, called a dipole distribution because it has two poles, implies that the earth has a velocity of about 300 kilometers per second as it moves toward the con-

stellation Virgo. Taking into account the motion of the solar system in the Milky Way galaxy, one can infer that our galactic center is moving at 600 kilometers per second relative to the CMB. This velocity is quite large (.2 percent of the speed of light), and its meaning, although not completely clear, may be related to large-scale matter flow in the universe.

Such flows are but one of the most recent conceptions that have revolutionized thinking about the evolution of the early universe. The hot big-bang model, the central idea a decade ago, is still a good general hypothesis, but it has been modified and expanded to encompass new empirical data and new concepts. One of the fundamental difficulties with the big-bang model is the "horizon problem." It arises in the explanation of how the temperature of the CMB can be uniform on large angular scales. Thermal equilibrium is established by the exchange of energy. The big-bang model requires that the early expansion be so rapid that regions of the sky now separated by more than two degrees in angle could never have exchanged energy with each other, even if the energy traveled at the speed of light. Why then do all the parts of the sky now appear to have the same temperature? The probability of this having

been a random occurrence is unimaginably small. A refinement of the model, a concept called the inflationary universe hypothesis, addresses the problem.

The inflationary picture explains the uniformity of temperature by postulating that the universe expanded at an enormous rate in its first few instants. In this model, the universe evolved from a small space that had been in thermal equilibrium prior to the time of inflation. Hence, the CMB is smooth because the CMB photons we see today issued from regions that were once in nearly perfect thermal equilibrium. What we see is therefore only a very small part of a region that had reached local equilibrium before the inflation.

A second consequence of inflation is that the density of the universe is required to be a unique value, called the critical density. The critical density is far larger than the observed density, which implies that the universe contains much more matter than just the luminous stars and galaxies that we observe. The unseen mass could be in the form of cold, dark matter that emits no detectable radiation.

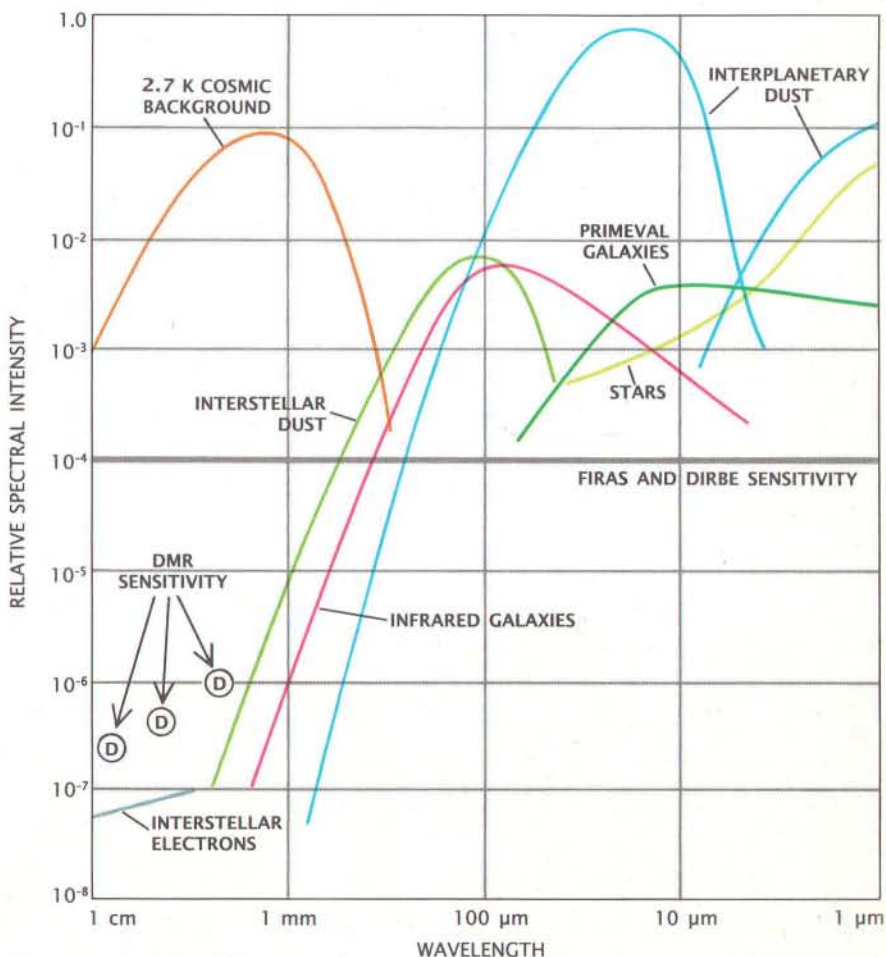
Inflation would also expand the quantum-mechanical density fluctuations that existed in the primordial material before the expansion began. This kind of fluctuation ought to have produced a characteristic form of anisotropy. Therefore, a third consequence of inflation is a prediction of how the CMB brightness should vary with angular separation across the sky. The discovery of fluctuations in the cosmic microwave background would be an important test of the validity of the inflationary model.

The extreme uniformity of matter in the early universe, inferred from the isotropy of the CMB, must be reconciled with the condensed structures we see in the present-day sky: galaxies, clusters of galaxies and superclusters. The dark-matter hypothesis has gained considerable favor in explaining this apparent dilemma. Incorporation of cold, dark matter into the big-bang picture has two effects. First, it permits the evolution of density perturbations to have started before the era of decoupling. This early formation of structure does not show up in the CMB, because the dark matter does not interact with the radiation. Second, the dark matter enhances the density in the universe and speeds the evolution of structure.

Studies of the distribution and redshifts of galaxies have led to surprising conclusions about the density and distribution of luminous matter. Enormous regions of space appear to be virtually free of any galaxies. These regions, called voids, are much larger than one would expect for randomly positioned matter that has clumped gravitationally. The theoretical challenge is to modify the hot big-bang picture so that one can reconcile the smooth distribution of the radiation at decoupling, represented by the cosmic microwave background, with the existence of large voids in the distribution of luminous matter. Examples of such modifications to the "standard hot big bang" cover a wide range of possibilities. Some cosmologists postulate that very small increases in the density cause very large increases in the tendency to form galaxies. Others propose that once cold, dark matter has clumped, it can decay, leading to

explosions on a truly cosmic scale. Such cataclysms would account for the observed voids.

Measured distributions of the velocities of galaxies imply the existence of very large-scale structures in the cosmos. The velocity distribution measurements are exceedingly difficult and fraught with systematic bias, but several independent studies indicate that something unexpected is occurring. Widely separated galaxies seem to be moving in unison at velocities that would not be expected from a standard model of the evolution of the matter. It appears that our galaxy is on the edge of an enormous region of space that is moving as though gravitationally drawn toward an object dramatically named the Great Attractor. The surprisingly large velocity of the Milky Way relative to the CMB, as measured by the dipole variation, may be related to this motion. Both the voids and the large-scale motions constitute



EXPECTED BRIGHTNESS of the phenomena COBE will observe is graphed as a function of wavelength. Local sources of varying brightness across the sky are plotted at their expected minimum values; estimated sensitivities of the COBE experiments are also indicated. The 2.7-K cosmic microwave background dominates local sources over a large range of wavelengths. Radiation from the early galaxies should be nearly as intense as emissions from interplanetary dust at wavelengths near five micrometers.

experimental evidence that the big-bang model, even with the addition of cold, dark matter, has difficulty encompassing.

The question of how structure evolved from the uniformity of the early universe is also addressed by measuring the properties of the radiation emitted after the decoupling era. If matter had been as luminous in early epochs as it is today, a measurable cosmic infrared background should have been produced. There are several pieces of circumstantial evidence for the presence of such early luminosity. The oldest stars observed are known to contain significant amounts of elements heavier than helium; such elements can only be formed by nuclear fusion in the interior of stars. Because the heavy elements produced in stellar interiors do not become detectable until they are ejected into the interstellar medium by a star's death, the observed heavy elements in the oldest

known stars must have been created in an earlier generation of stars.

It is likely, therefore, that the oldest stars we see are from at least the second generation. Where is the light from the first generation of stars? Were these first stars smoothly distributed in space or clumped together to form the first galaxies? If they were clumped, the radiation might be found in highly redshifted protogalaxies (whose spectra have been shifted to longer wavelengths by their sources' rapid recession from the earth's frame of reference). Such protogalaxies would appear as extended dim patches in the sky. If the stars formed earlier than larger-scale structures did, then their light would appear as a uniform glow in all directions.

It is possible that, after the production of the heavy elements, some of the ejected material precipitated as

dust. This dust would have absorbed the radiation from the first-generation stars. The dust would have heated up and reradiated the energy at longer wavelengths. The effect of such a process would be to make the energy from the early heavy-element-producing stars appear in the far infrared. If the density of the dust had been high enough, the angular structure of the infrared background would show noticeable variations.

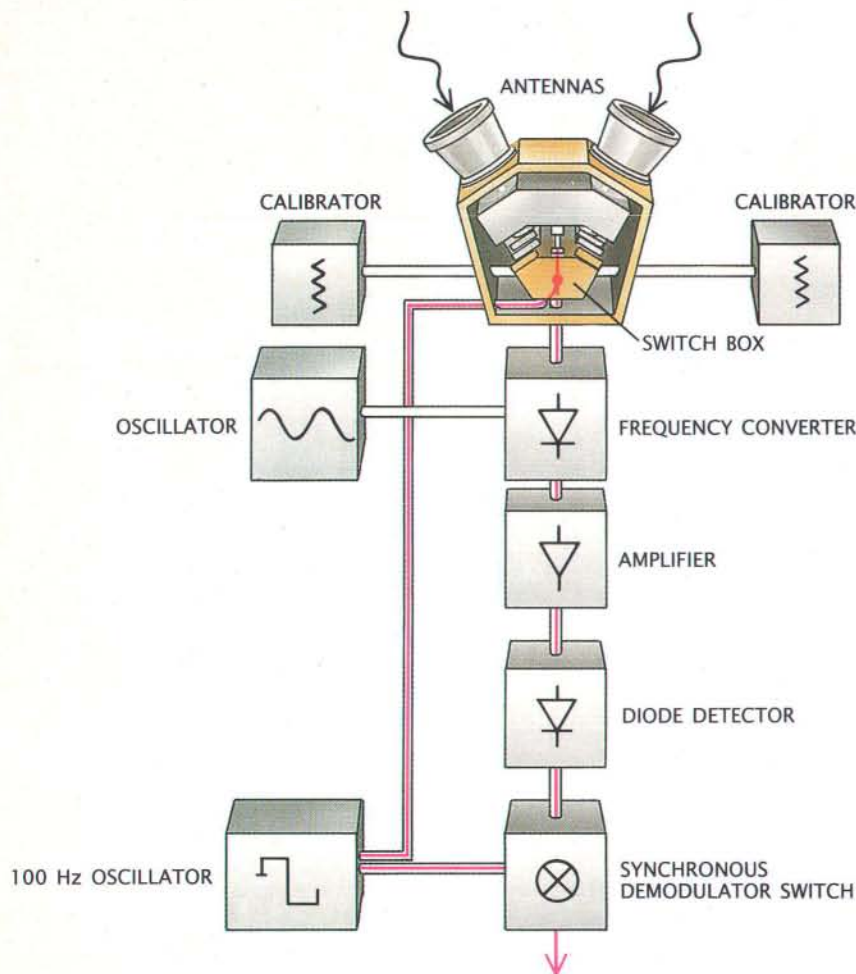
It is evident that many new and exciting ideas about the formation and early development of the universe have been raised and that comprehensive and detailed observations of the cosmic background radiation are the key elements needed for sorting out these ideas. The faintness of the cosmic background relative to the astrophysical and terrestrial radiation conspire to make ground-, aircraft-, balloon- and rocket-based experiments extremely difficult.

Properly shielded from the sun and the earth and oriented to provide full-sky coverage, a satellite, however, can measure a broad range of wavelengths. The satellite can thus deepen understanding of the cosmic background as well as measure radiation from local sources.

This local radiation, rather than instrument sensitivity, sets the ultimate limits on the ability to measure the cosmic background. The foreground astrophysical sources include dust in our solar system, synchrotron radiation from electrons losing energy in galactic magnetic fields, thermal radiation from interstellar dust in our own galaxy and the integrated emission from the stars and external galaxies. Because the various sources can be distinguished according to their spatial and spectral characteristics, it is possible to separate the foreground sources from the cosmic backgrounds. Well-calibrated, multifrequency, full-sky maps are required to perform this separation.

The COBE satellite was designed and built at NASA's Goddard Space Flight Center. The design combines a careful integration of instruments, spacecraft and orbit to reduce systematic errors, with instruments that cover a broad spectral range (near infrared to centimeter wavelengths) and can measure the background radiation across the sky.

Primary mission objectives are to search for angular anisotropies in the CMB, to measure its spectrum and to search for and measure the diffuse



DIFFERENTIAL MICROWAVE RADIOMETER measures the difference between the microwave radiation emitted from two points on the sky with two horn antennas that are alternately connected to a single receiver. Each horn's signal is compared with the signal from the other horn. This technique minimizes variations in receiver gain, variations that would reduce the sensitivity of systems using two receivers.

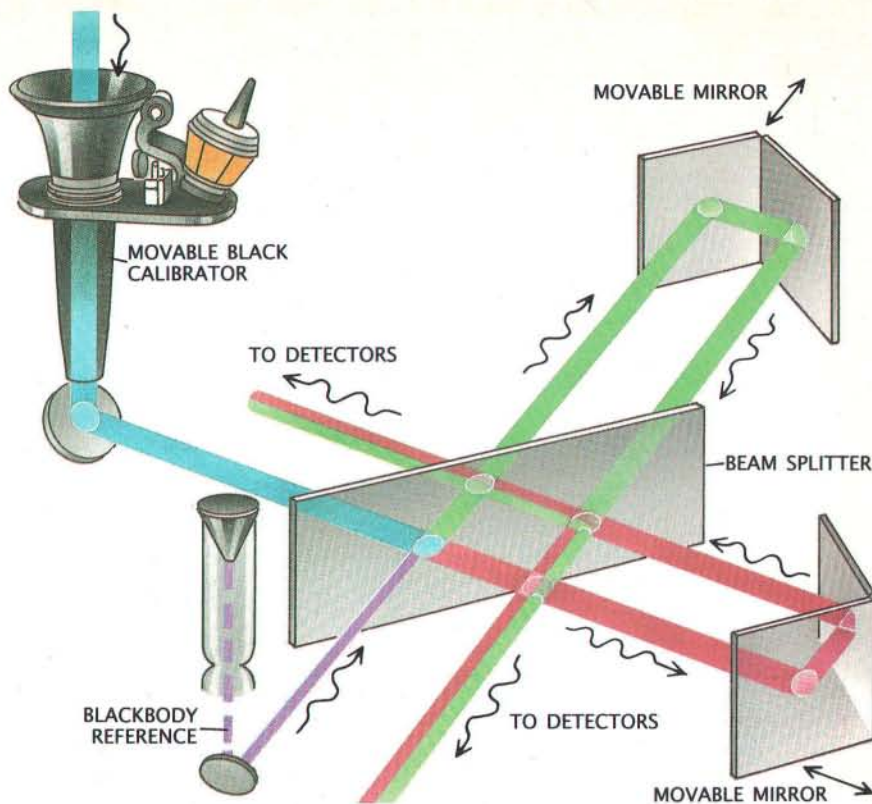
CIB. Analysis of emissions from the foreground astrophysical sources is also an objective of the mission. Such knowledge is intrinsically important to understanding the origin and evolution of both our solar system and galaxy. Furthermore, the cosmic backgrounds cannot be determined without this knowledge.

COBE carries three complementary detectors: a set of differential microwave radiometers, a polarizing Michelson interferometer and an infrared filter photometer. Each measures a different aspect of the cosmic background radiation. The instruments have each been designed to minimize systematic errors and to provide the sensitivity needed to measure the cosmic backgrounds.

The differential microwave radiometers (DMR's) will measure the large-scale anisotropy in the CMB to better than one part in 100,000, an extremely small variation. Data from this instrument will be used to search for the seeds of the present large-scale structures, anisotropic expansion or rotation of the universe, gravity waves, large-scale flows of matter and cosmic strings. Theory predicts that such strings—which can be considered as massive, one-dimensional objects—will have formed just after the big bang. If they did, they may have provided a framework on which large-scale structures could grow.

The radiometers achieve their high sensitivity by rapidly switching between two nearly identical horn antennas, each aimed at an angle of 30 degrees with respect to the spacecraft's axis of spin. Measured power differences are then converted to temperature differences in the sky by comparison with references supplied by onboard sources and measurements of the moon. There are three separate receiver boxes, one for each of the three wavelengths. The specific wavelengths were chosen to optimize the capability to distinguish between local galactic and dust emissions and the cosmic microwave background.

To provide higher sensitivity, critical components of the two shorter wavelength radiometers are passively cooled by radiation to about 140 K, while the longest wavelength radiometers operate near room temperature. The cooled radiometers can detect a temperature difference of about .025 K in a one-second measurement. The room-temperature radiometers are less sensitive by a factor of approximately two. The sensitivity of the data from each DMR receiver after one year



FAR INFRARED ABSOLUTE SPECTROPHOTOMETER compares the spectrum of radiation from the sky at wavelengths from 100 micrometers to one centimeter with that from an internal blackbody, or perfect source of radiation; differences between the two would constitute evidence that the universe underwent energetic processes near the time of decoupling. A trumpet-shaped horn funnels light into a beam splitter, which directs two parts (red and green lines) along paths whose lengths are varied by movable mirrors. The components are then recombined at the beam splitter to form an interference pattern that reveals the signal's spectral nature.

of observation will be about 100 microkelvins (.0001 K) per seven-degree field of view on the sky. This level of sensitivity is about seven times better than that of the three-millimeter map shown on page 124.

By combining these points, the DMR will be able to measure temperature variations on large angular scales as small as 10 microkelvins. Such variations are 300 times smaller than the amplitude of the confirmed dipole variation mentioned earlier. Even the dipole temperature difference caused by the motion of the earth around the sun with respect to the CMB will appear as a large signal relative to the instrument's noise.

The Michelson interferometric spectrometer, or far infrared absolute spectrophotometer (FIRAS), will measure the spectrum of the background radiation from one centimeter to 100 micrometers for each of 1,000 parts of the sky. Deviations from the spectrum of a blackbody will be measured to within one part

in 1,000. A deviation would indicate the presence of very energetic sources in the early universe. Scattering of cosmic background photons by hot electrons produces a well-known change in the blackbody spectrum. This perturbation would indicate the presence of a hot, ionized gas produced by energy injection well after decoupling. Such heating could have been produced by the formation of stars or galaxies.

The FIRAS, like the DMR, is a differential instrument. It compares the spectral power received from the sky with an internal reference source that has a controllable temperature and calibrated emission properties. The high accuracy of the FIRAS instrument is attributable to a movable calibration source that may be placed in the entrance of the input horn. The spectrum emitted by the calibrator is within .01 percent of a blackbody. The temperature of the calibrator is adjusted to match the flux from the sky as closely as possible. Any remaining spectral differences between the blackbody and the sky

can then be measured at a high degree of sensitivity.

Radiation from the sky enters the instrument through a trumpet-shaped cone that suppresses the off-axis radiation. The resulting beam is split into two components, which traverse paths

whose lengths are controlled by mobile mirrors. The spectrum is inferred from the way the waves of the two beams interfere with one another after they recombine. The instrument's field of view is seven degrees, directed along the spacecraft's spin axis.

The third detector, the diffuse infrared background experiment (DIRBE), will measure the absolute brightness of the sky at wavelengths ranging between one and 300 micrometers. This instrument will perform the most sensitive search yet undertaken for the diffuse infrared light from the early universe—light from the first generation of protogalaxies, galaxies and stars. The spectral range of the radiation would indicate the nature of its originating processes. DIRBE will also make important measurements of the emissions from foreground sources such as interstellar and interplanetary dust, galactic starlight, infrared galaxies, quasars and galactic clusters.

DIRBE's optics have been designed and built to eliminate all stray light from off-axis sources as well as radiation from the spacecraft and the instrument itself. A system of light baffles, radiation stops and extremely clean, highly polished mirrors ensures that the radiation contributed by extraneous sources will be small. The instrument will be able to measure a small residual background radiation with a sensitivity of about 1 percent of the local astrophysical foreground emission. The basic instrument design is that of an unobscured off-axis Gregorian telescope with a primary mirror diameter of 20 centimeters; the field of view is .7 by .7 degree. The telescope is aimed 30 degrees away from the satellite spin axis so that the rotation varies the angle between the DIRBE line of sight and the sun. A device resembling a tuning fork interrupts the sky beam 32 times per second to compare the incoming radiation from each point on the sky with the near-zero light level of a cold reference surface within the instrument. Detectors at 10 different wavelengths observe the same field simultaneously to cover the spectrum from one to 300 micrometers.

The DIRBE will also measure the polarization of the incoming light in the three shortest-wavelength bands. This information will enable it to characterize the sunlight that is scattered from the interplanetary dust. The four middle bands are dominated by the thermal emission of the interplanetary dust, but the receivers for those bands are sensitive enough to detect the diffuse emission from an early generation of stars. The channels measuring the longest wavelengths will search for radiation that may have been reradiated by intergalactic dust produced from this early generation of stars. DIRBE has enough spectral coverage and sensitivity to separate

INSTRUMENT PARAMETERS AND SCIENCE TEAM MEMBERS

PARAMETER	DIFFUSE INFRARED MICROWAVE EXPERIMENT (DIRBE)	DIFFERENTIAL MICROWAVE RADIOMETERS (DMR's)	FAR INFRARED ABSOLUTE SPECTROPHOTOMETER (FIRAS)
WAVELENGTH	1.1–1.4 15–30 2.0–2.4 40–80 3.0–4.0 80–120 4.5–5.1 120–200 8.0–15 200–300 (in μm)	3.3, 5.7 and 9.6 mm	.10–10 mm
SPECTRAL RESOLUTION	Shown above	1 GHz	.2 cm^{-1}
FIELD OF VIEW	.7 degree square	7 degrees diameter	7 degrees diameter
TYPE	Filter photometer	Dicke-switched differential radiometers	Fourier-transform polarizing interferometer
FLUX COLLECTOR	Off-axis Gregorian telescope	Corrugated horns separated by 60 degrees	Smooth flared horn
SENSITIVITY (IN FIELD OF VIEW)	$10^{-13} \text{ W cm}^{-2}$.0001 K (3.3 and 5.7 mm), .00025 K (9.6 mm) (after 1 year)	$< 10^{-13} \text{ W cm}^{-2}$ (.5–5 mm)
LINE OF SIGHT	30 degrees off spin axis	30 degrees off spin axis	Spin axis
DETECTOR	Photovoltaics for < 5.1, photoconductors for 8–120, bolometers for 120–300	Diode mixer	Bolometers (4)
CALIBRATION	Internal reference and celestial sources	Noise diode and moon	Blackbody temperature controlled to within .001 K
DATA RATE (BITS/SEC)	1,713	216	1,330

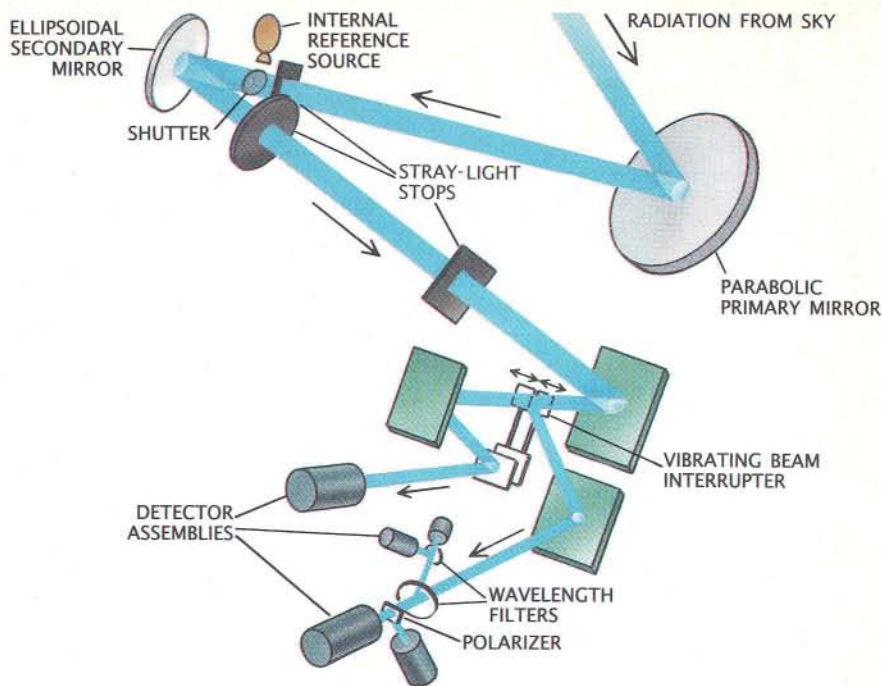
COBE Science Team Members

Goddard Space Flight Center:	Charles L. Bennett, Nancy W. Boggess, Eli Dwek, Michael G. Hauser (principal investigator for DIRBE instrument), Thomas Kelsall, John C. Mather (COBE project scientist and principal investigator for FIRAS instrument), S. Harvey Moseley, Jr., Richard A. Shafer, Robert F. Silverberg
Massachusetts Institute of Technology:	Edward S. Cheng, Stephan S. Meyer, Rainer Weiss (chair of the COBE science team)
Jet Propulsion Laboratory:	Samuel Gulkis, Michael A. Janssen
University of California, Santa Barbara:	Philip M. Lubin
General Research Corporation:	Thomas L. Murdock
University of California, Berkeley:	George F. Smoot (principal investigator for DMR instrument)
Princeton University:	David T. Wilkinson
University of California, Los Angeles:	Edward L. Wright

the emissions of planetary and galactic dust from the interesting cosmic sources. The questions of whether radiation was emitted from an early generation of stars or larger structures and whether such radiation might have been absorbed and reemitted by intergalactic dust are uniquely suited to an analysis based on DIRBE's observations.

Among the satellite's other components are a large cryostat (a vacuum-insulated tank containing liquid helium), a deployable radiation shield, a power system and an attitude-control system. The shield protects the sensitive instruments and the cryostat from solar and terrestrial thermal radiation and from radio-frequency interference. The DIRBE and FIRAS are mounted inside the cryostat to maintain them at a temperature of less than 2 K; this arrangement minimizes radiation from the instruments themselves and permits the use of sensitive detectors. The cryostat contains enough superfluid helium to chill the instruments for the nominal mission lifetime of about one year, the time needed to achieve the required sensitivity and full-sky coverage. The DMR is mounted outside the cryostat, but it, too, is protected by the radiation shield. Solar-cell panels supply electric power to the satellite except during short eclipse periods, which occur during only part of the year. At such times batteries will be used.

The entire satellite, comparable in size and weight to a large automobile, was lifted into orbit by a Delta vehicle, launched from the Western Space and Missile Center in California. The COBE orbit allows the three kinds of scientific instruments to scan the entire sky while keeping them in a stable thermal environment with minimal interference from the sun and the earth. It is a nearly polar, circular orbit at an altitude of 900 kilometers, arranged to remain near the earth's day-night terminator (the border of the sunlit portion of the globe). The attitude-control system keeps the instrument's view directions aligned about 90 degrees from the sun and 180 degrees from the earth. The entire spacecraft rotates at .8 revolution per minute, producing a scan pattern that reduces systematic errors in the DMR and giving DIRBE a range of solar-illumination angles from which to view reflections from the interplanetary dust. Because the brightness of the interplanetary dust depends strongly on the ecliptic latitude and the sun's illumination angle, the emission from these particles



DIFFUSE INFRARED BACKGROUND EXPERIMENT searches for the radiation from the earliest generation of stars, scouring it for clues to the ancient distribution of matter from which today's cosmic structures evolved. Light is collected by a primary mirror; stray radiation is eliminated by light stops and baffles. The beam is divided into 10 components, which pass through different wavelength filters. The bands are analyzed for their intensity; three are analyzed for their polarization properties.

will be easy to recognize and thus easy to remove from the data. The rotation also ensures that the sun heats the satellite uniformly, reducing temperature gradients within the satellite.

The data gathered by COBE's instruments will constitute a set of fundamental information of unprecedented scope and accuracy. The information will be analyzed and plotted as maps of the whole sky that span four orders of magnitude in wavelength. Two sets of results will be published and delivered to the National Space Science Data Center within three years of the launch. The first is a set of maps, calibrated and corrected for all known instrumental and spacecraft effects. The second set will show the microwave and infrared backgrounds that remain after the effects of the local astrophysical sources are removed.

The data from COBE will answer many questions about the early universe. Some of these have been asked for centuries while others have arisen more recently, as a result of new evidence. We can anticipate that cosmology will make a leap forward because of COBE discoveries.

In the long run, however, it is the comprehensive data set itself that will be COBE's greatest contribution. That set will be vastly more valuable than the sum of its parts. The uniformity

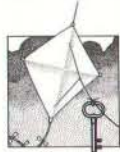
of the analysis, the ability of each instrument to confirm the results of the others and the sheer completeness of the information will lead to a level of reliability unachievable in a mission of lesser proportion. Future investigations of the early universe and large-scale structure, in whatever direction the field progresses, will depend on COBE's legacy.

FURTHER READING

- THE BIG BANG. Second Revised and Updated Edition. Joseph Silk. W. H. Freeman & Co., 1988.
- THE COSMIC BACKGROUND RADIATION AND THE NEW AETHER DRIFT. Richard A. Muller in *Scientific American*, Vol. 238, No. 5, pages 64-74; May, 1978.
- THE LARGE-SCALE STRUCTURE OF THE UNIVERSE. Joseph Silk, Alexander S. Szalay and Yakov B. Zel'dovich in *Scientific American*, Vol. 249, No. 4, pages 56-64; October, 1983.
- THE INFLATIONARY UNIVERSE. Alan H. Guth and Paul J. Steinhardt in *Scientific American*, Vol. 250, No. 5, pages 90-102; May, 1984.
- VERY LARGE STRUCTURES IN THE UNIVERSE. Jack O. Burns in *Scientific American*, Vol. 255, No. 1, pages 30-39; July, 1986.
- DARK MATTER IN THE UNIVERSE. Lawrence M. Krauss in *Scientific American*, Vol. 255, No. 6, pages 50-60; December, 1986.

THE AMATEUR SCIENTIST

A backyard version of a Stirling engine can be built with common materials



by Earl Walker

Heat engines, which convert heat into useful mechanical work, are of two broad types: those in which combustion operates directly on a piston and those in which it operates indirectly by way of an intermediary known as the working fluid. The first type is an internal-combustion engine, of which the gasoline engine is the obvious example: when fuel is burned, the gaseous combustion products expand directly against a piston. The second type is an external-combustion engine. One example is the steam engine, in which water is the working fluid. First a fuel—coal, say—vaporizes the water; then the steam is introduced into a cylinder and expands against a piston.

Another example of an external-combustion engine is one that was introduced in Scotland in 1816 by the Reverend Robert Stirling. Originally its working fluid was air; later designs have used hydrogen or helium. The Stirling engine is interesting for several reasons. It recycles its working fluid continuously. Any source of heat will do, so that a fuel can be chosen for its low level of pollution. And at least in theory it should be highly efficient in converting heat into work. Nevertheless, for a variety of reasons Stirling's idea lost out—first to steam and then to internal combustion.

Recently the idea has made something of a comeback, in part because of the low-pollution possibility and the fact that the fuel need not be petroleum-derived. The engine has also caught the attention of some amateur scientists. One such tinkerer has been Peter L. Taler of the Windfarm Museum on Martha's Vineyard, Mass., who modeled an engine after one developed in 1876 by A. K. Rider of Philadelphia. Taler's apparatus is unlikely to compete with conventional engines because its output power is low. Still,

it is easy to construct from common materials and allows one to study the associated thermodynamics.

One appealing feature of the apparatus is that it does not require finely machined cylinders and pistons, as other Stirling engines and all internal-combustion engines do. Instead it uses two cans (soda cans, say), which are partially submerged in water [see illustration on opposite page]. The water is contained in two tanks at the base of the apparatus. Each can is attached to the end of a rod; the other end of each rod is connected to a crank on a weighted flywheel at the top of the apparatus. An air-filled tube runs between the tanks and up into the can in each tank.

When the water in one of the tanks is heated by some source, such as a flame, the air in the interconnecting tube shuttles back and forth between the tanks, the cans rise and fall and the flywheel turns at several tens of revolutions per minute. These motions depend on two subtle features of the apparatus. One feature is the arrangement of the cranks at the flywheel: their outer arms are perpendicular to each other (as seen from the side). The other feature has to do with the way heat is transferred from the source to the air in one of the cans.

Before I explain the details of Taler's apparatus, I shall examine the basic principles of a Stirling engine with a textbook version, which is illustrated at the bottom left on page 132. Two solid pistons fit snugly within a cylinder and can be moved to the left or right either by air pressure inside the cylinder or by machinery to which they are attached. At the center of the cylinder there is a porous material, such as a metal mesh, called the regenerator, which temporarily stores heat when the engine is running. Near the pistons there are two "thermal

reservoirs," where the temperature is kept constant: on the left a "heat reservoir," maintained at a high temperature by a heat source, and on the right a "cold reservoir," whose temperature is kept low by some means of heat drainage.

During the engine's operation, the internal air undergoes cyclic variations of pressure, temperature and volume: the air is said to change in state. The piston arrangements for four of the states are shown in the illustration. The associated variations in the state of the air are best understood by following the graph of pressure versus volume at the bottom right on page 132. During its operation, the engine effectively cycles clockwise around a skewed and distorted rectangle on the graph.

First consider state 1, which corresponds to the first of the series of drawings and also to the upper left corner of the skewed rectangle. Piston *B* on the right is adjacent to the regenerator; piston *A* on the left is somewhat farther from the regenerator. The air trapped between the pistons is at a high pressure. As the heat reservoir warms the air, the air expands, thereby pushing *A* to the left; the consequent increase in the volume of the space between the pistons diminishes the pressure. During the expansion, the temperature of the air is kept constant because of the proximity of the heat reservoir, and so the expansion is said to be isothermal. The expansion is represented by the upper curve on the graph's skewed rectangle. When *A* reaches its leftmost position, the air is in state 2.

Next both pistons are moved to the right—not by heating but by the machinery to which they are attached—until *A* reaches the regenerator and *B* is all the way to the right. The air is then in state 3. The movement of the pistons causes the air to flow through the regenerator, which takes up some of the heat and thereby cools the air. Because the pistons move in synchrony, the volume of the air does not change during this transition, and so the transition is said to be one of constant volume.

Now the machinery attached to *B* pushes the piston toward the left. As the air is compressed, it gives up heat to the cold reservoir. Because the reservoir is fixed in temperature, the temperature of the air does not change and the transition is said to be an isothermal compression. At the end of the compression, the air is in state 4. To complete the cycle, the machinery moves both pistons together to the

left until they are again in the arrangement for state 1. Again the transition takes place at constant volume. When the air flows through the regenerator, it regains the heat it lost in the previous constant-volume transition.

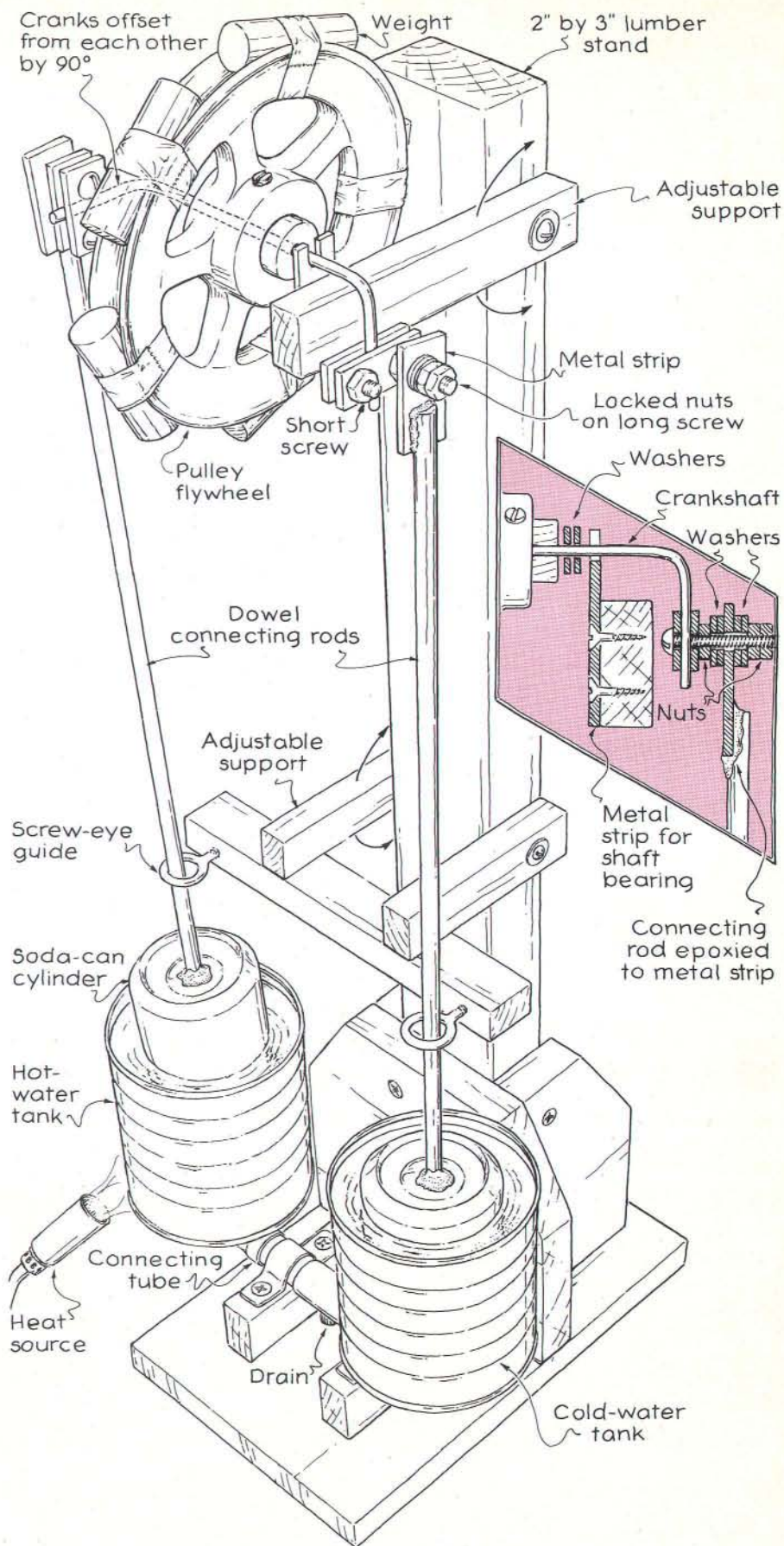
As the engine runs, it continues to cycle around the closed curve on the graph. In the transition from state 1 to state 2, one of the pistons is moved by expansion of the air. In the other three transitions, the pistons are moved by the machinery. Is the engine useful, that is, does the air do more work on the machinery than the machinery does on the air?

To answer the question, I must first explain what is meant by "work." Although commonly the term can mean almost any expenditure of energy, it has a narrower scientific definition: work is the transfer of energy that is needed to move something. On that restricted definition, if a force is applied to an object but the object does not move, no work is done. If there is movement, then the amount of work is the product of the force and the object's displacement. Whatever provides the force loses energy, and the energy shows up as motion.

Suppose that there is a closed container of air and that the air and the container are at the same temperature. The air molecules beat incessantly against the walls of the container, and the collective outward force on a wall from the collisions is the air pressure against that wall. If the wall does not move outward because of the pressure, the air performs no work on the wall. But if the wall yields, work is performed. For air to do work, then, it must somehow expand its container. In principle the work is the product of the force in each molecular collision and the displacement of the wall caused by that collision. The work is more easily expressed, however, as the product of the pressure (which is analogous to force) and the change in volume (which is analogous to displacement).

If an external force were to shrink the container, work would be done by that force rather than by the air. The agent responsible for the force (which might be you or some machinery) would then lose energy, which would be transferred first to the wall and then to the air molecules as the wall moved inward. In this case, work is done against the air. The amount of work is again the product of the pressure and the change in volume.

The idea behind an air-filled Stirling engine is to coax the air into doing work against a piston—pushing the



Peter Tait's Stirling-engine apparatus

piston outward and increasing the volume of the space between the pistons. The motion of the piston can then be transferred to machinery, where the acquired energy can be put to use. If there were only one such expansion, of course, the engine would hardly be helpful. The engine must instead somehow compress the air periodically so that the air can expand periodically and thereby continue to do work. In short, the volume of the air must be cycled repeatedly. But remember that for the air to be compressed, the machinery must do work on the air. If the machinery does as much work on the air in the course of a cycle as the air does on the machinery, the engine produces no net work and is useless.

The solution to the problem involves the temperature of the air. Suppose that whenever the air does work, it is hot. Then the collisions of the air molecules on the piston are vigorous, and the pressure is high. Because the work done on the piston depends directly on the pressure, the amount of work is large. Next suppose that whenever the machinery does work on the air, the temperature is low. Then the collisions are weaker, so too is the pressure, and the amount of work done on the air is small. If the temperature can be adjusted in this way, the air does more work on the machinery than the machinery does on the air.

Such a periodic variation in temperature and pressure lies behind the textbook Stirling engine (as well as other engines, in fact). Work is done by the air on piston A during the isothermal expansion, when the air temperature is high. Work is done by the machinery on the air during the isothermal compression, when the temperature is low. The engine has a net output of work.

The work involved in a cycle of the engine can be derived from the graph of pressure versus volume. During

the isothermal expansion, the amount of work done by the air is represented by the area below the corresponding curve. The area is bounded by the curve, the volume axis of the graph and two vertical lines that extend from that axis up through the end points of the curve. During the isothermal compression, the amount of work done on the air is the area below the corresponding curve. No work is done during the constant-volume transitions because there is no change in volume, and so the area below those lines on the graph is zero. To find the net work done by the engine during a full cycle, you subtract the area beneath the compression curve from the area beneath the expansion curve. The result is the area within the skewed rectangle.

I now return to Tailer's apparatus. The heated tank is the heat reservoir. The other tank is the cold reservoir, whose temperature is maintained by thermal radiation and convection. The air-filled spaces, including the tube that connects the tanks, serve as the cylinder. Either the tube itself or some wire mesh that can be placed inside it functions as the regenerator. The machinery to which the cans are attached is the flywheel.

The series of drawings at the top of page 134 indicates how the air trapped within the lower section of the apparatus responds to the heating and to the motion of the flywheel. The drawings show the elevation of the cans and the water, the direction of air flow and the orientations of the cranks for eight stages. The labels on the cranks indicate whether they are connected to the hot or the cold tank.

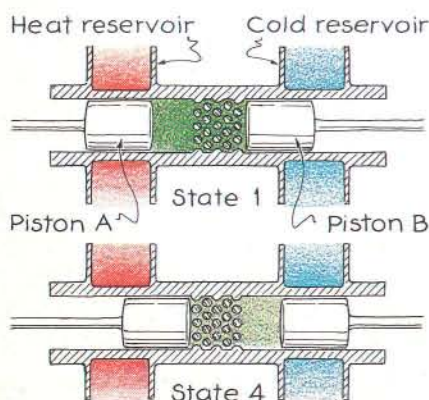
Tailer's engine is similar to the textbook engine but lacks any true isothermal or constant-volume transitions. Still, if you were to graph air pressure versus air temperature, the engine would cycle around on the graph somewhat as I described for

the textbook engine. To follow the cycling, consider the engine as it goes through stage *a*, having just left stage *h*. During *a* the hot can rises faster than the cold one sinks. Next both cans rise until they reach *c*. Then the cold can rises faster than the hot one sinks, until they reach *d*. Notice that during the transition from *h* to *d* there is more air in the hot can than in the cold one. This means that more of the air is being heated than is being cooled, and so the air pressure increases. Notice too that during the transition from *h* to *d* the volume of the air increases. The expansion is driven by the additional pressure, which means that the air does work on the cans—and thus also on the flywheel.

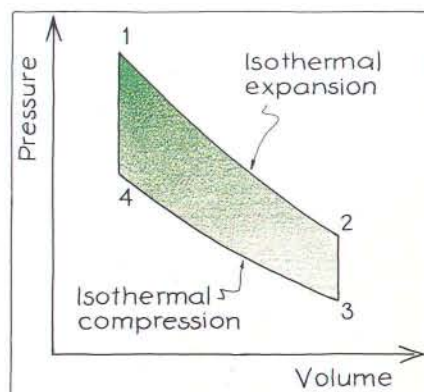
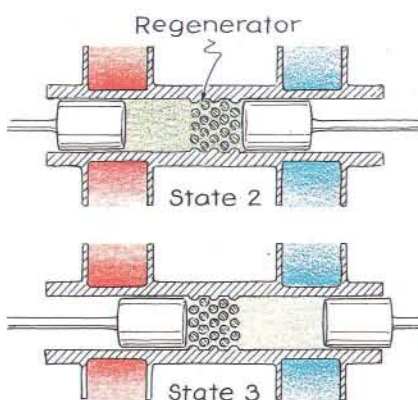
When the engine moves between *d* and *h*, the variations in volume and pressure are just the reverse, and the flywheel does work on the air. The net motion of the cans compresses the air; the shift of air to the cold can diminishes the overall heating of the air and decreases the air pressure. During compression, the pressure is low, and so the work done by the flywheel on the air is less than the work that was done by the air on the flywheel during the earlier, *h*-to-*d* transition. The result is a net output of work by the air.

Tailer sent along specific plans for constructing his engine, but he points out that the details are easily varied according to the materials available. Fashion the crankshaft from stiff wire, such as sturdy clothes-hanger wire, so that it does not flex during the engine's operation. The crankshaft rests on aluminum strips 1/8 inch thick that serve as bearings. Drill holes through each strip, cut a notch at the top to support the crankshaft and then screw the strip to the interior of a wooden arm as shown in the illustration on the preceding page.

The flywheel is a pulley eight inches



Four states of a textbook Stirling engine



The pressure-volume cycle

in diameter with a groove designed to accept a V-shaped belt. Its bore hole is fitted with a short length of wood through which a central hole is drilled for the crankshaft. Glue the wire to the wood insert with epoxy, and secure the insert to the pulley with the set screw on the pulley. Bring the crankshaft wire out past the bearings, and then bend and cut the ends so that about two inches of wire extends perpendicularly from the crankshaft axis. The end sections should also be perpendicular to each other as seen from the side.

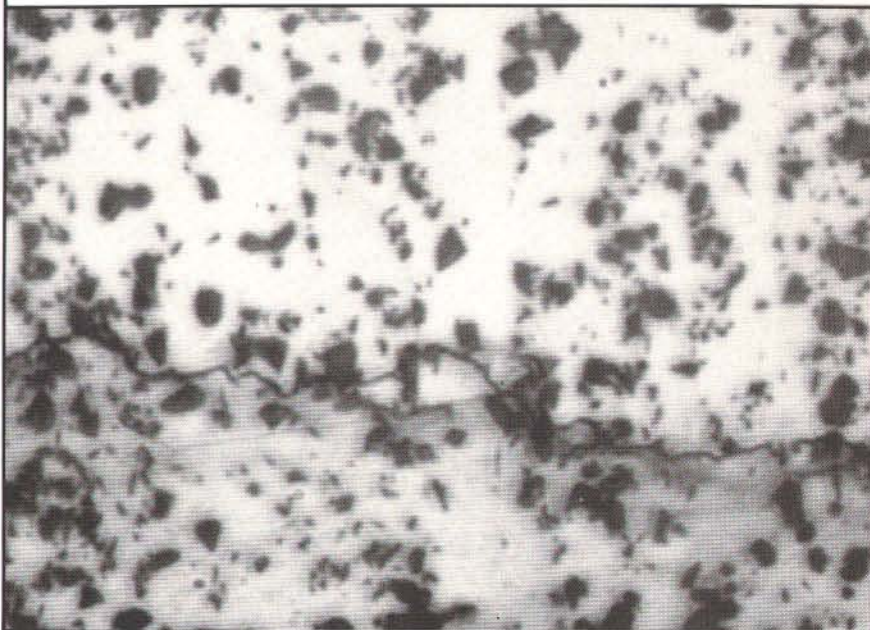
Cut the tops off two soda cans with a hacksaw. Invert each can and glue a wooden dowel to the bottom of the can. These connecting rods should be about 3/16 inch in diameter and about 36 inches long. For the glue Taiter suggests the type of epoxy that requires slow curing; it withstands heat better than the fast-curing type. Later, during final assembly of the engine, the upper end of each connecting rod is glued or taped to a strip of aluminum 1/8 inch thick that is mounted on a long machine screw with nuts and washers. The screw also passes through two other aluminum strips that are fastened to the outer ends of the cranks with a shorter machine screw and a nut. The entire assembly is called a crank journal.

The tanks are one-pound coffee cans. The tube linking them is made of sections of copper tubing whose internal diameter is 3/4 inch. Before you connect them, you should attach a section 5 1/2 inches long to each can. Make radial cuts in the bottom surface of a can with a knife, and force the tube section inward through the flaps left by the cuts, leaving about an inch of tubing below the can.

You can seal the tubing to the can with slow-curing epoxy, but soldering works better. If you choose to solder, sand the surface of the tube to be soldered, smear soldering flux on the surface and on the adjacent region of the can, and direct a torch onto the solder so that it flows over the flux. Once both cans are prepared, lay them on their open ends. Then sand and coat with flux the interiors of the tubing elbows, and solder the elbows in place. Next add two short lengths of tubing and a central coupling that is equipped with a drain. Solder them together except for the piece that inserts into the elbow at the cold tank. Secure that last connection with a few tight turns of vinyl tape so that the assembly can later be disconnected if a regenerator is to be added or replaced. If you cannot locate a tubing

QUESTAR® QM-100

unequalled long-distance microscope



High-resolution video image of a metallic composite sample at six inches working distance, field of view 175 microns. The fatigue crack is shown in a matrix of particles, the larger of which are typically 7-10 microns in width. Bright-field illumination is integral to the optic.

Questar Corporation is proud to announce the QM-100, the latest in its series of long-distance microscopes. It distinguishes 1.1 micron objects clearly at 6 inches, 1.5 microns at 10 inches, provides more than 1,500 times magnification on a twelve-inch video monitor, with matched and aspherized optics hand-crafted in the United States.

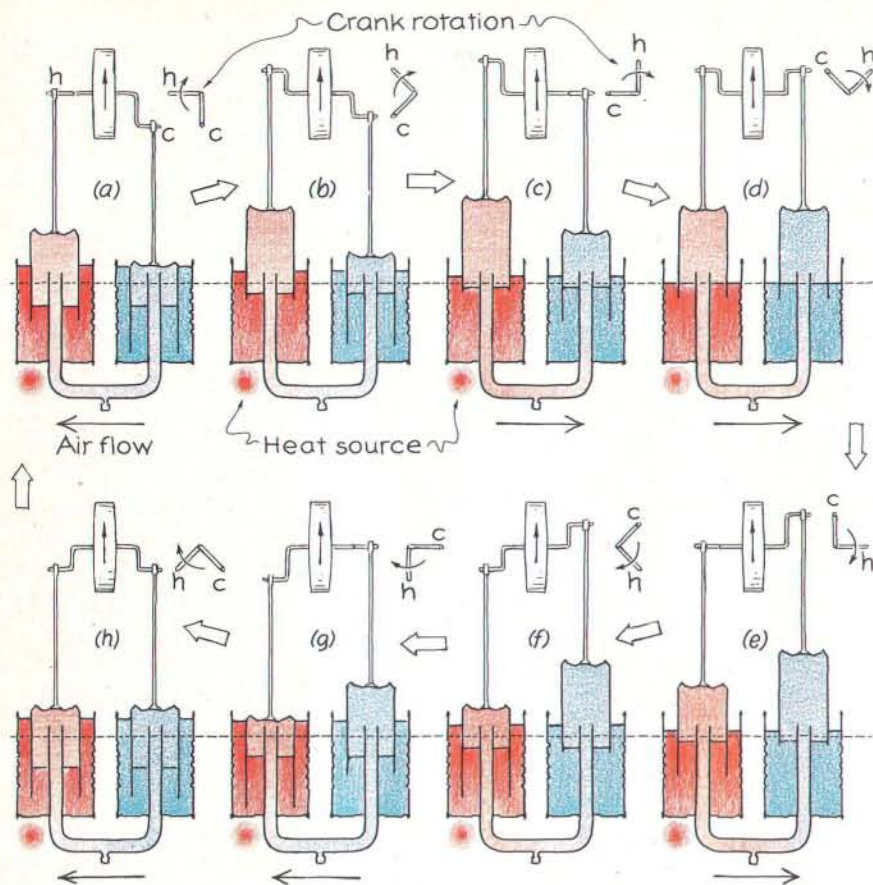
Since 1954 Questar's superb telescopes have also been used as long-distance microscopes. (Indeed at a distance of 108 inches they are still unsurpassed for special micro uses.) In 1983 we developed the QM-1, which identifies 2.5 micron objects clearly at 22 inches; it received an IR-100 award as one of the major technological achievements of that year. The QM-100 is then the newest member of a unique family.

At any distance — six, twenty-two, or sixty inches — no other instruments can match our optical performance, simplicity of use, and sheer technical perfection. For all your laboratory and production applications where extreme resolution is needed, call or write us today.

The Questar Long-Distance Microscopes
QM-1, QM-2, DR-1, QM-100

QUESTAR

P.O. Box 59, Dept. 221, New Hope, PA 18938
215-862-5277 • FAX 215-862-0512



Eight stages of operation of the Tailer apparatus

section with a drain, simply drill a hole in a regular tube, smooth the sides of the hole and then close it with a wood or rubber plug. Construct a cradle that will support the tank assembly and that allows easy access to the drain.

Now run the dowels through screws on the support column, and at-

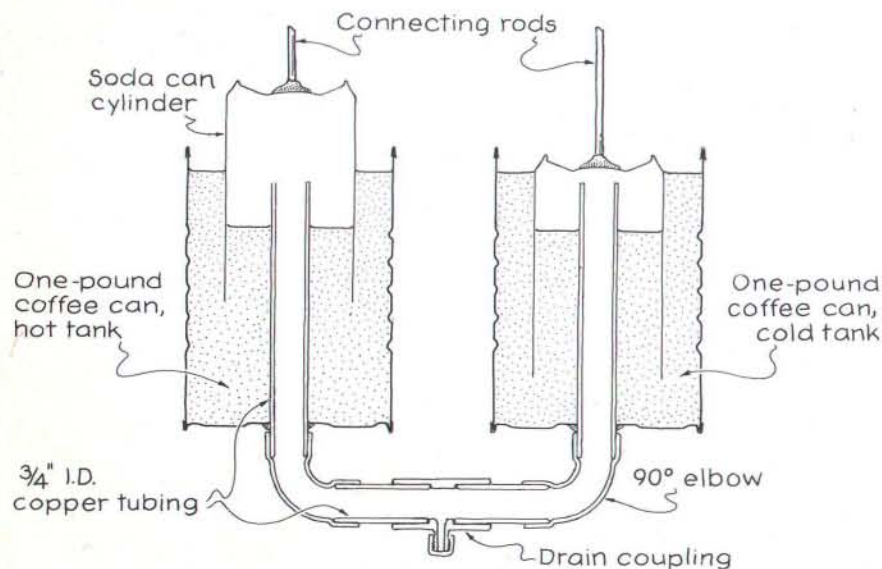
tach them to the outer metal strips of the crank journals as described above. To get a stroke length of 1.5 inches, set the screws on the journals so the long screw moves $\frac{3}{4}$ inch below the shaft and $\frac{3}{4}$ inch above it when the flywheel turns. Tape weights such as machine bolts to the flywheel to give it

enough mass to complete a rotation when the engine is operated. Then oil the bearings and make sure that the flywheel and cans move easily.

To ready the engine, turn the crankshaft until both cranks are pointed up at 45 degrees to the vertical. Then, with the drain open, fill the cold tank with cold water until it overflows into the interconnecting tube and goes out through the central drain. Next pour hot water into the hot tank until it too overflows. Close the drain and begin to heat the hot tank with, say, a propane torch or Bunsen burner.

The speed at which the flywheel turns depends on the temperature difference between the two tanks. For example, one of Tailer's engines ran at 20 revolutions per minute when the water temperatures were 200 degrees and 60 degrees Fahrenheit but speeded up to 28 revolutions per minute when the hot water was brought closer to boiling. The operation of the engine can be enhanced if the interconnecting tube is partially filled with rolled strips of wire mesh to act as a regenerator. When Tailer added several such rolls to his engine, it rotated almost once per second.

In addition to varying the temperature, you might try adjusting several other parameters of Tailer's apparatus. If the stroke length is varied, does the flywheel turn faster? What happens if the angle between the cranks is varied somewhat from the 90 degrees I have described? (Indeed, why does the angle matter, and why should the hot crank lead the cold crank?) Can other types of regenerator material improve the engine? Does performance increase if you substitute another liquid for the water? (Do not use any liquid that might result in a fire or explosion!) What happens if you alter the length of the connecting rods to increase or decrease the average height of the column of air in the cans?



The tube-and-can assembly

FURTHER READING

- THE STIRLING ENGINE. Graham Walker in *Scientific American*, Vol. 229, No. 2, pages 80-87; August, 1973.
- STIRLING ENGINES. Graham Walker. Oxford University Press, 1980.
- LIQUID PISTON STIRLING ENGINES. C. D. West. Van Nostrand Reinhold Company, 1983.
- OTHER EXTERNALLY REVERSIBLE CYCLES. J. B. Jones and G. A. Hawkins in *Engineering Thermodynamics: An Introductory Textbook*. John Wiley & Sons, 1986.
- PRINCIPLES AND APPLICATIONS OF STIRLING ENGINES. Colin D. West. Van Nostrand Reinhold Company, 1986.

YOU HAVE TO BE BLIND TO GET ON THIS PLANE.

Or you have to be an eye surgeon.

That's because Project ORBIS is a flying eye hospital. And its mission is to train doctors while treating patients throughout the world.

Last year alone, ORBIS helped over 1000 blind men, women and children regain their sight.

It also helped more than 1000 surgeons learn and use the latest ophthalmic techniques. Techniques that restore sight or save it.

That's important. Important because at least thirty million blind people and 350 million other people with potentially blinding diseases can be cured.

That's right, they can be cured. All of them. They just need more doctors with more training.

And ORBIS provides that training to those doctors. Right on board its aircraft.

What's more, as ORBIS spreads its treatment and training, it spreads goodwill. People regain their faith as they regain their sight.

So give people a better vision of the world. Send your tax-deductible contribution to Project ORBIS, Suite 1900, 330 West 42nd St., New York, NY 10036. **Project ORBIS**



COMPUTER RECREATIONS

The cellular automata programs that create wireworld, rugworld and other diversions



by A. K. Dewdney

"The chess board is the world, the pieces are the phenomena of the universe, the rules of the game are what we call the laws of Nature."

—THOMAS HENRY HUXLEY,
A Liberal Education

If a chessboard represents the world, so does a cellular automaton. Its gridwork of squares will support many more pieces than chess, and its possible rules are myriad. Even better, its colors are not restricted to black and white but span the spectrum. What is a cellular automaton? As far as this month's examples are concerned, it is an infinite two-dimensional plane filled with squares, a collection of states and a clock. Each square or cell is always in one or another of the available states. At each tick of the clock each cell changes its state in accordance with certain rules that apply in its neighborhood.

Two software packages that gen-

erate cellular automata have recently caught my eye: the PHANTOM FISH TANK and RUDY RUCKER'S CA LAB. People who have no special knowledge of programming can now explore certain famous cellular automata, or they can create cellular worlds of their own.

Over the years this department has described many cellular automata. In October of 1970 the famous cellular automaton called life, which was discovered by mathematician John Horton Conway of the University of Cambridge, was first celebrated in these pages by Martin Gardner. Life is full of curious structures—some stable, some moving. In strange ways the growth of these structures mimics the development of a colony of bacteria. The August, 1988, column presented a cellular automaton that simulated a self-sustaining chemical reaction. More recently, this past August, we witnessed the evolution of a cellular automaton called cyclic space from a

state of random disorder into a patchwork of beautiful crystalline spirals competing for space.

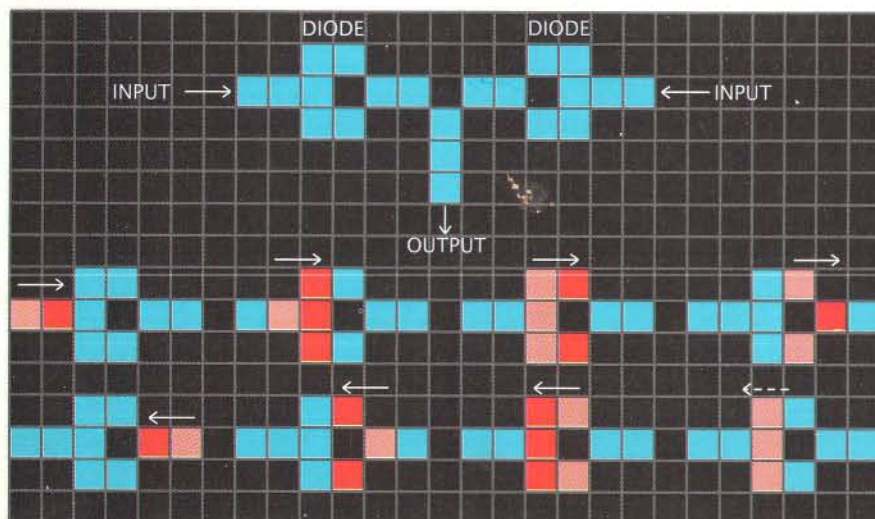
Each of these cellular automata requires different sets of rules that govern the way the states flit about among the cells. Traditionally, to experiment with a new cellular automaton, one had to develop a new program embodying its rules. Wouldn't it be nice to have a single system that somehow embodied all possible cellular automata?

Such a thought certainly motivated Brian Silverman to produce the PHANTOM FISH TANK in 1987. (Silverman works as a research director at Logo Computer Systems, a Montreal software firm; he was one of the moving spirits behind the Tinkertoy computer described here this past October.) The PHANTOM FISH TANK refers whimsically to the computer screen as a tank and to the strange patterns that sometimes writhe or glide in cellular space as fish. My own favorite cellular automaton in Silverman's package allows one to build a simple computer within a two-dimensional cellular space. I call it wireworld.

With the editing system available in the PHANTOM FISH TANK, one can design and animate circuits of cellular "wires" and "logic gates." In theory these cellular devices can be assembled into a computer of any power. In practice the scale of the wireworld computer is limited to a small number of gates and wires.

A handful of devices is more than enough, however, to impress any avid experimenter with the limitless possibilities of wireworld. The building blocks of wireworld are an array of square cells. At any given moment a cell can be in any one of four possible states. The states have names instead of numbers: background, wire, electron tail and electron head. A pair of adjacent cells, an electron head and tail, make up an "electron."

It hardly matters what one calls such cells, of course, but the names have significance when it comes to the rules. During each turn of the game electron-head cells become electron-tail cells, electron-tail cells become wire cells and sometimes wire cells become electron-head cells; background cells never change. Whether a wire cell changes depends on the state of its neighbors, the eight cells that touch it along an edge or at a corner. Wire cells whose neighborhood includes one or two electron-head cells become electron-head cells themselves, but wire cells whose neighbor-



OR gate is protected by diodes (top). An electron passes through a diode the right way (upper sequence) but is blocked going the wrong way (lower sequence)

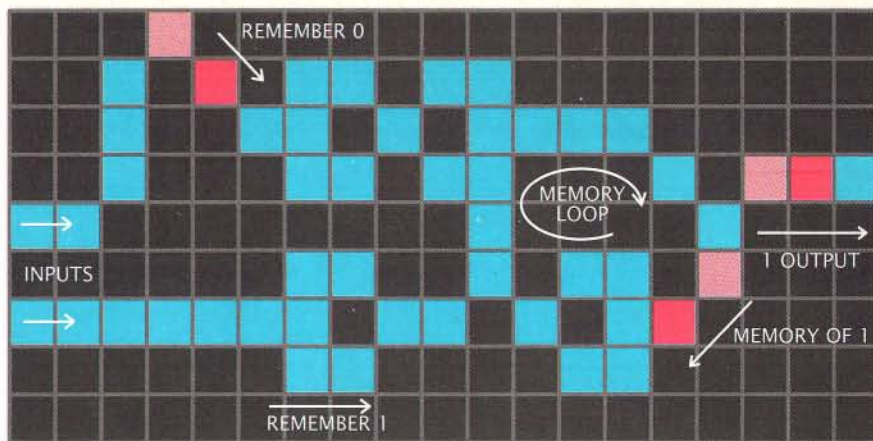
hood is crowded by three or more electron-head cells remain wire cells.

With rules no more complicated than that, wireworld is ready for action. A random distribution of electron cells and wire cells will probably do little more than short circuit, though. In order to compute anything, the human designer must first assemble logic devices out of the wire and the background cells. Then he or she can test the devices by introducing electrons that glide along wires that look like nothing on earth. Building the cellular computer is half the fun.

To make a wire in the cellular space of wireworld, one simply positions several wire cells in a line and surrounds them with background cells. To release an electron into the wire, one replaces the first wire cell in the line with an electron-tail cell and the second with a head cell. The electron will move along the wire as a simple consequence of the wireworld rules: at each tick of the cellular clock the wire cell in front of the electron becomes an electron head, even as the head cell becomes a tail cell and the tail cell reverts to wire.

But how are the decision-making elements fashioned for the wireworld computer? Consider the configuration in the upper part of the figure on the opposite page. The two horizontal wires are inputs for an OR gate, a device that will send an electron down the vertical wire if an electron enters either input wire or both of them. At first glance the gate might seem to work without the strange rectangular knobs placed on the input wires: if either of the input wires carries an electron or if both of them do, at least one will arrive at the junction, and one electron will certainly enter the output wire. The problem occurs when a single electron arrives at the junction. It would not only continue down the output wire but also propagate a copy of itself along the other input wire. The situation would be disastrous if another electron entered the device along the same wire and collided with the stray electron. They would cancel each other.

The rectangular knobs on the input wires prevent such collisions. They are cellular diodes, which allow electrons to travel through the circuit in only one direction. The diode consists of a three-by-two rectangle made mostly of wire cells. The central cell on one side of the rectangle is a background cell, creating a gap in the wire. Nothing more is needed, as the diagram on the opposite page shows. An electron that



Memory element in wireworld is about to forget 1 and remember 0

enters from the solid side will split in two and then trigger the wire cell on the other side of the gap. An electron that goes the opposite way splits into three electrons that cannot propagate down the wire, because the wire cell adjacent to them cannot be triggered by three electrons. The OR gate is easily constructed by placing a diode on each input wire.

Two more kinds of gates will complete the repertoire at the component level: a memory element and a NOT gate (or inverter). If one attempts to design a complete computer, one would certainly have to decide on a timing convention for internal signals. In ordinary electronic computers the binary signal does not consist of individual electrons but of voltages that are constant, either "high" (1) or "low" (0). In wireworld the presence or absence of an electron will be interpreted as 1 or 0, respectively. But computational events must nevertheless be orchestrated within the cellular computer. The individual electron signals must be spaced out in time. A constant number of clock cycles will separate consecutive signals along any wire inside our computer. If the constant number of clock cycles is T , then for every T cycles every component of the computer will be receiving either a single electron (1) or no electron (0). The question is, How small a value of T can we get away with?

I will leave the enjoyable business of designing an inverter to readers after supplying a few hints. Set up a loop of wire cells around which a single electron can circulate every T seconds. Tap into the loop by a wire that leads to a configuration resembling a backward diode. A wire from the outside world brings an electron (or nonelectron) every T ticks. If an electron arrives, the electron from the loop enters the qua-

si diode in time to cancel it. If no electron arrives, the electron from the loop escapes into the output wire.

In logical terms wireworld can now be equipped with a complete stock of components. Inverters and OR gates suffice to build any logical function whatsoever, from multiplexers (circuits that direct the flow of signals within a computer) to a CPU (the central-processing unit—a circuit of mind-boggling complexity that runs a computer's programs). This column will not attempt such things; it will be content merely to show that everything one needs to build a computer is already available.

But what of memory? How shall we build a computer's memory registers? The standard recipe calls for a so-called flip-flop made out of the logic that lies at hand. I find more charming the prospect of designing a custom memory element. The illustration above depicts the device I designed with help from the PHANTOM FISH TANK (to lay it out and test it).

The memory element employs a loop within which a single electron circulates when the element remembers a 1; when it remembers 0, no electron will be found there. Two control wires change the content of the memory element. An electron moving along the lower wire, signaling the new memory of a 1, enters the loop at its lower left corner. It makes no difference whether an electron is already in the loop. The entering electron is injected into the circuit precisely when a circulating electron would pass the injection point were it present.

The upper wire sends nonelectrons into the loop. An electron that enters the circuit along the upper wire signals that until further notice the element should remember 0. The electron passes through a protective di-

ode and enters another diode. The latter is actually part of the memory loop. The signal electron causes all three cells on the loop side of the diode to become electrons at exactly the moment when a circulating electron would be about to enter the diode from the bottom. This effectively kills the circulating electron by turning all the cells that were at one moment electron heads into electron tails. The latter are refractory and cannot become heads again for one whole cycle. The loop "forgets" the 1 and now remembers 0. Of course, if it were previously remembering 0, no electron would be circulating anyway. In that case an electron in the lower cell of the diode would propagate into the loop in the "upstream" direction. But not to worry: the stray impulse is blocked at two other diodes before any damage can be done. One of these is in the loop, and the other protects the lower input wire.

How long should the clock cycle be to ensure that all runs smoothly? The smallest memory loop I could manage takes 13 ticks. If no reader can devise a faster memory element that does everything mine does, we shall have to be content with a grand cycle of 13 ticks of the clock.

Silverman is enthusiastic about the computational potential of cellular automata. He imagines a cellular computer that not only operates on the basis of circuit layouts of the kind I have described but also can modify the circuits on the fly to optimize somehow the computation being carried out.

Rudy Rucker sees a still wider future for cellular automata. "I feel that science's greatest task in the late 20th century is to build living machines.... This is the computer scientist's Great Work as surely as the building of the Notre Dame cathedral... was the Great Work of the medieval artisan."

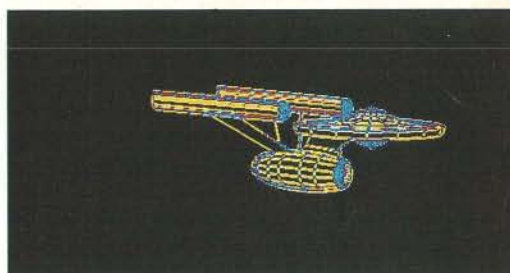
Whether that is a viable project or not, readers can certainly have fun trying, now that Rucker, in cooperation with Autodesk, a California computer-games company, has produced the package called RUDY RUCKER'S CA LAB. (Rucker is a science-fiction writer who left the uncertainties of a freelance life for the relative security, but great excitement, of academic life near California's Silicon Valley.) The LAB consists of two programs called RC and CA. The first program, written by Rucker, enables a beginner to experiment with a sampler of various interesting rules. These include those for the well-known cellular automata

such as life and the voting game [see "Computer Recreations," SCIENTIFIC AMERICAN, April, 1985] and a number of newer rules with names such as faders, ascii and Brian's brain (a cellular automaton first described by Silverman). The CA section, written by John Walker of Autodesk, includes a special facility to create almost any cellular automaton one likes.

One of Rucker's favorite cellular automata, called rug, creates patterns of color that are arranged in oval swatches and elliptic bands. The patterns look rather like the hooked rugs my ancestors used to make in Lancaster County, Pa.—but the cellular automata rugs are alive: the colors are always on the advance or retreat, and the pattern changes continually as we watch. How does Rucker's rugworld work?

In the rug cellular automaton, each cell has eight neighbors, four along the sides and four at the corners. A cell decides which of 256 states it will enter at the next cycle by invoking a four-stage process. First, the average of the eight neighboring states is computed. Second, the resulting average is truncated to an integer by removing the fractional part. Third, an increment (selected by the user) is added to this number. Finally, the resulting sum is masked with another number supplied by the user. "Masking" means that the bit representation of a cell's new state is combined logically with the bit representation of the user-supplied number. For example, if the numerical result of the computation just outlined happens to be 107 and the mask number is 224, then the masking process would compare the binary representation of both numbers: 107 equals 01101011 in binary; 224 equals 11100000. The mask produces a new binary number that has 1's only where both numbers have 1's. The result of the masking process in this example, then, would be 01100000. The mask in effect chops off the low-order bits of the computed number to obtain the next state. Of course, one can use the number 255 (11111111) as the mask, in which case the results of the computation are left intact and the next state of each cell is essentially the average of its neighbors incremented by the amount specified by the user. Even as small an increment as 1 produces patterns that tend to spread or shrink, keeping the rugs "alive."

For mental adventurers who wish to roam, the CA program allows them to specify any rule they like in a variety of modes and languages. One can specify the size of the cellular



space as well as the effects on a cell that crosses the boundaries. Will the cell jump across to the other side of the screen or drift away into computer memory?

For less serious players there is always the possibility of selecting one's favorite cellular automaton from a variety of demonstrations and then starting the automata from different initial configurations. It can be troublesome, however, to specify the initial state of 8,000 cells. Luckily, Rucker has provided the enjoyable option of using recognizable pictures as the starting configurations, such as the picture of the Starship Enterprise shown on these two pages. Its pixels are in reality the colored cells of a cellular automaton. Start off the rug cellular automaton from this initial configuration, and the Starship Enterprise weaves itself into a high-tech rug. Run a cellular automaton called heat, which simulates heat flow in materials, and the Enterprise glows red, violet and chartreuse as it diffuses slowly into the cosmos. (There is no chance for the crew to beam away.) Or start the life cellular automaton and watch the Starship Enterprise explode into a cloud of gliders, blinkers, beacons and beehives.

The PHANTOM FISH TANK for Apple II computers can be ordered by writing Brian Silverman at Logo Computer Systems, Inc., 3300 Cote Vertu, Montreal, Quebec, H4R 2B7. RUDY RUCKER'S CA LAB for IBM PC's can be ordered from Autodesk, Inc., 2320 Marinship Way, Sausalito, CA 94965.

This past June a parade of prose and poetry, or should one say, a parody of prose and poetry, appeared in this department. MARK V. SHANEY, a program that reads straightforward English text only to produce monstrously distorted versions of the same, works on the basis of Markov chains. The underlying algorithm is ridiculously simple. As the program scans the text, it builds up a table of



The Starship Enterprise dissolves in the cellular space called life

two-word followers. In other words, for every pair of consecutive words that the program finds in the text under examination, it lists all the words that follow that pair wherever it occurs throughout the text. Frequencies of such following words are easily converted into probabilities at the end of the process, so that as MARK V. SHANEY regurgitates the text, it merely looks up the last two words it printed, selects a follower based on the probabilities involved and then prints the follower as its next word. It repeats the process ad nauseam.

I shuddered at the thought of MARK V. SHANEY going through that very column, reducing my own carefully constructed text to gibberish. Kenneth A. Bullis of Campbell, Calif., incarnated SHANEY on his Macintosh Plus computer, then promptly, and with no thought for my personal feelings on the matter, typed in the entire "Computer Recreations" for June! Here is the program's first go at the column:

"Take care of a Markov chain table for all words that rhyme with sun: bun, done, fun, gun and so forth. Scanning them, one's eye might alight on the nature of a solid or a liquid. Look around. What are the only ones that take part in the leaves of green plants. It is only the Mandelbus' first stop can be obtained by writing Pike at the moment when the sad clowns enter your museum of pain."

Guy Ottewell of Greenville, S.C., thought that the poetry software did not perform as well as SHANEY. In particular, he took issue with the sonnet produced with the aid of Michael Newman's computer-assisted poetry program. The fault lay not in the program but in the poet. The program, by the way, runs on IBM computers and can be purchased in three parts: the POETRY PROCESSOR, NERD II (the acronym stands for Newman's Electronic Rhyming Dictionary) and ORPHEUS A-B-C. Alternatively, readers can order all three in one package called the POETRY PROCESSOR PACKAGE. The

programs are available through the *Paris Review*, 541 East 72 Street, New York, NY 10021.

Another rhyming dictionary is produced by Carl Wurtz of Burbank, Calif. QUICKRHYME, as Wurtz calls his program, has apparently tackled many of the thornier issues connected with computing rhymes rather than merely listing them. Readers interested in QUICKRHYME can contact Wurtz at "a priori," 859 Hollywood Way, Suite 401, Burbank, CA 91510.

In August, 1989, this department explored the cellular automaton called cyclic space by its discoverer David Griffeath of the University of Wisconsin at Madison. I can do little more than sample the interesting experiments being carried out as a consequence in the scientific hinterland by hundreds of dedicated amateurs.

Griffeath's cellular universe is called cyclic because of the "eating" habits of its cells. A cell in state k will eat a neighboring cell in state $k-1$. If k happens to be 0, the next lower state is $n-1$ if only n states are allowed. The food chain becomes a food cycle. If the cellular space is initially filled with random states (debris), when n is large, there is little activity at first, because it is rather unlikely that any particular cell would be adjacent to a cell in the next lower state.

Harry J. Smith of Saratoga, Calif., has worked with a 320-by-200 space where each cell is in one of 14 states. At first about 25 percent of the cells change their states at each iteration, but by 15 generations the percentage drops to less than three. Then, Smith says, the patches of intense activity that Griffeath calls droplets begin to form. As the droplets grow, the percentage of cells that change state at each iteration grows steadily until, at about the 150th generation, the percentage begins to grow even more rapidly until the demon phase is reached, when all of the cells are changing at every iteration.

In the column on cyclic space I chal-

lenged readers to discover a small rectangle of cells that do not contain a defect but will in time produce one. By a defect Griffeath means a closed cellular chain that includes a contiguous cycle of all possible states. Such is the egg from which demons hatch. The nicest solution to this problem comes from Marlin Eller of Seattle, Wash. What happens to the following three-by-three rectangle in an eight-state cyclic universe?

```
0 1 2
7 4 3
7 5 6
```

John D. Brereton of West Haven, Conn., developed an ingenious screen display to show what the cyclic cellular automaton was really up to. He produced not one picture of cyclic space on his screen but two, side by side. The space on the left shows the last generation, but the space on the right shows only the cells that change their state as the next generation is computed. Unable to leave things as Griffeath and I left them, Brereton tinkered by counting as neighbors not only the four cells along a side but also those touching a cell's corners. The new space prompted the feeling of discovery: "As we approach this strange planet, at first we see a multicolored cloud cover, beneath which no details are visible. As we go closer, descending through successive cycles, color patterns on the ground are vaguely distinguishable through gaps in these clouds. Then distinct boundaries of various colored fields appear, followed by small buildings..." But I leave the description to be completed by readers who wish to visit Brereton's planet.

FURTHER READING
CELLULAR AUTOMATA MACHINES. Tommaso Toffoli and Norman Margolus. The MIT Press, 1988.

Authors... LOOKING FOR A PUBLISHER?

**Learn how to have
your book published.**

You are invited to send for a free illustrated guidebook which explains how your book can be published, promoted

and marketed. Whether your subject is fiction, non-fiction or poetry, scientific, scholarly, specialized, (even controversial) this handsome 40-page brochure will show you how to arrange for prompt publication.

**To the
author
in search
of a
publisher**



Unpublished authors, especially, will find this booklet valuable and informative. For your free copy, write to:
VANTAGE PRESS, Inc. Dept. F-53
516 W. 34 St., New York, N.Y. 10001

SCIENTIFIC AMERICAN

**is now available
to the blind and
physically handi-
capped on cassette
tapes.**

All inquiries should be made directly to RECORDED PERIODICALS, Division of Associated Services for the Blind, 919 Walnut Street, 8th Floor, Philadelphia, PA 19107.

ONLY the blind or handicapped should apply for this service. There is a nominal charge.

BOOKS

*Two roads to the stars, art of the people,
continuity and catastrophe*



by Philip Morrison

NORTON'S 2000.0: STAR ATLAS AND REFERENCE HANDBOOK, edited by Ian Ridpath. Eighteenth edition. Co-published by Longman Scientific & Technical and John Wiley & Sons, Inc., 1989 (\$34.95). **DO-IT-YOURSELF ASTRONOMY**, by Sydney G. Brewer. Edinburgh University Press, 1988. Distributed by Columbia University Press (paperbound, \$15).

At this perihelion we look to sky events a little farther ahead than usual. The star atlas was first compiled by the English amateur astronomer Arthur P. Norton in 1910, to grow edition by edition into the most widely used such atlas in the world. It neatly maps the entire starry sky visible to the unaided eye in satisfying form, one page for each polar cap and six two-page spreads of "shield-shaped gores" that march around the celestial equator. The virtue of Norton's map design is that the pages have ample room for detailed entry, and yet each spread spans enough sky to make big star patterns clear. This new computer-aided version plots 8,700 stars, down to the faintest to be seen in velvet-dark skies, marking among them the Milky Way, double stars, some 500 variable stars and 600-plus deep-sky objects, as well as all the star clusters, nebulae and galaxies accessible to binoculars or small telescopes. The map is up-to-date enough to include the supernova SN 1987A. The coordinate grid is that of the millennial year, as the title promises.

With its indexes and the full tables that catalogue all "interesting objects" map by map, the atlas itself covers about 40 pages. The other three quarters of the handy volume has reference materials on astronomy: some 60 numerical tables and lists with definitions, constants and articles of practical aid to the observer, newly prepared by a long list of specialists, amateur and professional. This edition retains the comfortably moderate level of technicality of its predecessors. The

geometric framework, the complexities of astronomical time measures, the physical data of planets and satellites are set out here not diagrammatically but numerically.

This is a rocky path for sheer beginners, but the compact, explicit and crisp style makes Norton a quick and comprehensive reference for anyone who is past that stage and is interested in the positional and optical astronomy of brighter objects. Galaxies are touched on only slightly.

You will have to look up in some other annual almanac the positions of planets during any particular year. Here there is a longer view: for instance, the rare occasions when all four bright moons of Jupiter happen to be hidden from the telescope. (The next time happens to be June 15 of this year, although only early in evening twilight for American watchers.) Eclipses are well heralded; many Americans will enjoy a lengthy total solar eclipse in the summer of 1991. For finding even a faintly visible star in the sky, identifying a long list of stars by name, naming moon craters or Mars markings, observing Mercury, plotting a sunspot path, grasping the apparent rotation of the moon, watching for meteors or an aurora, selecting a small telescope or looking up a big star catalogue—for all of these the pages here are just right.

The second book is most original. Sydney Brewer is a retired college mathematics teacher and an amateur astronomer near Edinburgh. Gifted with a deft sense for the essentials of inference and a flair for apt instrumental design, he set himself the task of "simply going out under the stars to rediscover from the most basic measurements as much astronomy as possible." Within a couple of years he was able to gauge the cosmos a long ways out, honestly linking a steel one-foot rule there on his desk with the distances to the stars. His clear book, with only a little mathematics, will

carry along many a delighted armchair collaborator.

Brewer used no telescope, took no long voyage, employed nothing beyond the resources of a domestic workbench except two grand gifts of our times: a familiar quartz watch with the radio means to keep it accurate to the second day after day and a small calculator, its internal algorithms in silicon, a device that confers on any of us the power to outreckon even tireless Kepler.

His first goal was nothing less than the primum mobile itself. How fast does the sky turn? This is living-room astronomy, good for city dwellers and comfortable even on winter nights. Find a window from which you can view any bright star that in its night path appears from or hides behind any fixed obstacle: some neighbor's wall or chimney acting as a distant front sight. Watch through your windowpane, where for a rear sight you glue a small square of black paper with a pupil-size viewing hole. Time the star's appearance or disappearance in that peephole to the nearest second or two over an interval of a couple of months, counting the days carefully. You will have seen "at first hand the silent precision with which the star appears or disappears, always at the same spot and always on time." Two long runs gave him the sidereal day good to a part in a million.

Well armed with the exact period of the first moving thing, he then caught on a squared paper fixed below a skylight a pinhole image of the afternoon sun, marking its position at the appointed second over a few days. Month after month he repeated the sun measurement; between one autumn and the next spring he could plot the changing angular motion of the sun as compared with the unseen stars. The orbit of the earth was his to admire. Here he assumed the equal-areas law of Kepler to find the earth ellipse itself, although only in relative terms.

There is no way to fix the solar-system scale in length units without fine telescopic angular measures, distant voyages, rare events or radar; recall the 18th-century efforts to observe the transits of Venus. Brewer fishes rough guesstimates by physical argument out of simple brightness comparisons among the cloudy or rocky planets. He makes that method work better still over the light-years that part us from the nearest stars. All the work is clearly set out with simple theory and real data.

The radius of the earth itself is

To preserve your copies of SCIENTIFIC AMERICAN

A choice of handsome and durable library files or binders. Both styles bound in dark green library fabric stamped in gold leaf.

Files Each holds 12 issues. Price per file \$7.95
• three for \$21.95 • six for \$39.95

Binders Each holds 12 issues. Issues open flat.
Price per binder \$9.95 • three for \$27.95 •
six for \$52.95

(Add \$1.00 per unit for postage and handling in the U.S. Add \$2.50 per unit outside the U.S.)

**To: Jesse Jones Industries, Dept. SA, 499 East Erie Ave.,
Philadelphia, PA 19134**



file

binder

Send me _____ SCIENTIFIC AMERICAN

☐ Files ☐ Binders

For issues dated ☐ **through 1982** ☐ **1983 or later.**

I enclose my check or money order for \$ _____;
(U.S. funds only).

☐ **Charge my credit card \$ _____ (Minimum \$15)**

☐ Amex ☐ VISA ☐ MasterCard ☐ Diners

Card # _____ Exp. date _____

Signature _____

or call Toll Free: 1-800-972-5858

Name _____
(please print)

Address _____
(No P.O. Box please)

City _____

State _____ Zip _____

SATISFACTION GUARANTEED. Pennsylvania residents add 6% sales tax.

measured by a modest journey that scales our planet in terms of a small steel rule, at three steps remove, to astonishing three-figure accuracy. What he measures is the shift in the time of day observed for the image of the declining sun to reach a neatly marked place in a well-leveled instrument, as the clever little device is itself shifted by a calibrated bicycle ride for several hours eastward on the highway. Brewer's demonstrative tour de force ends with an attic Foucault pendulum that turns with the sky, somehow aware of a remote dynamic background. (The deep reason is probably not gravity, as Brewer—after Mach—proposes, but the simple kinematics at great distances forced by long-past cosmic inflation.) Surely some readers are ambitious enough to embark on such a trip as Brewer's along their own private path.

THE SPIRIT OF FOLK ART: THE GIRARD COLLECTION AT THE MUSEUM OF INTERNATIONAL FOLK ART, by Henry Glassie. Color photographs by Michel Monteaux. Harry N. Abrams, Inc., 1989 (\$60).

A polychrome procession crowds in from the left edge of the endpapers to spill its devils, elephants, bandsmen, burros, market women, pilgrims and ritual dancers across five pages, announcing in proper style the marvelous book at hand. You are not yet past the overture; after the contents list, a second gathering appears, one more contained but with equal magic. The door of a little old church stands open behind the happy mother who is holding her baby at the baptismal font. Through the open door we see the sunny courtyard thronged with friends in their finery, who crowd the square to the green facade at the other side. Every one of these figures is six inches high, made of clay or straw or fabric, given its form in a score of lands around the world. About 280 more big color plates follow, along with 50 or so pictures in black-and-white that show a full-size world, mainly the artists in their workshops, men, women and family groups.

This entire cornucopia of folk images tumbles from a still larger source, the unique assembly of 100,000 objects from 100 countries that is the Girard Collection, established at the International Museum of Folk Art in Santa Fe during the early 1980's under the auspices of the Museum of New Mexico. It was on a visit to Mexico in the 1930's that Alexander and Susan Girard first became intrigued by the folk art they saw there in wood,

terra-cotta and papier-mâché, a lively world remade in small. Sandro Girard is an architect-designer, a colorist and an organizer extraordinaire, who has created many celebrated interiors around a rich theme, none more striking than this lifelong collection.

The Girards understood at once that folk art does not come by ones; it is often multiple, like the people themselves and their animals and their houses, and the art will reward considered attention to that central fact. They went on to collect objects everywhere by the dozens and the fifties. The vignettes they assembled, as ordered and as tangled as life itself, these miniature theaters and village squares, merry festivals and hushed ceremonies, are the heart of the collection.

This is not at all to say that the single pieces—the figurative dolls in costume or puppet masks or the abstract, patterned Acoma pots, geometric fabrics, painted gourds—are of any less importance or beauty. They too are here, all photographed with deceptive ease by the lens artist Michel Monteaux. Consider one spread among hundreds: in wood from Oaxaca, Emiliano Zapata in bright-blue uniform on his horse, a small Lisbon rider in glazed earthenware and a Yoruba horseman of heroic stature on his small mount, also in painted wood. That array of masterly single pieces strikes a second grand chord: a deep unity somehow flows from the cosmopolitan diversity of such small objects.

Henry Glassie, the distinguished folklorist at Indiana University who chose this sample from the enormous Girard Collection, has contributed a poetic essay that examines the meaning and nature of folk art, less for any taxonomy of styles than for the aesthetic and philosophical insights that dwell in the material as a whole. Much of his account draws on sympathetic encounters with the artists themselves, for most of the pieces are contemporary. One lifelong riverboatman in western Pennsylvania makes toy boats: not models but sculpture. Looking at the scale models in the museum, he said: "Hell, anybody can make a boat that looks like a boat." The real deck rails are small, but if you are in a storm on the river, you want them thick, so he makes them thick. Museum models look like riverboats, but the things Lou Seshier makes *feel* like boats. A conscious artist, he is at the private extreme of a scale that at the other end touches traditional public ritual.

There is no place for a full account

of Glassie's rich and sensitive distinctions, but his closing pages offer a summary eloquence. "Folk' art exists only because 'fine' art does.... We begin... by understanding each tradition in the purity of its own system.... Knowing the world... we will come to know art as mixed, as a message about the wonders of impurity.... Not fine, nor folk, nor primitive, not sensual nor conceptual, useless nor useful, traditional nor original.... Art is the best that can be done." These pages glowingly document the high claim for those of us who are far from Santa Fe.

CATASTROPHIC EPISODES IN EARTH HISTORY, by Claude C. Albritton, Jr. Chapman and Hall, 1989 (\$29.95). **THIS DYNAMIC PLANET: WORLD MAP OF VOLCANOES, EARTHQUAKES, AND PLATE TECTONICS**, compiled by Tom Simkin, Robert I. Tilling, James N. Taggart, William J. Jones and Henry Spall. Smithsonian Institution and U.S. Geological Survey, 1989. USGS Map Distribution, Federal Center, Box 25286, Denver, CO 80225 (\$4 postpaid, paper wall map).

"In the early 1930's when I took my first course in geology," Professor Albritton begins, "I was taught... the 'principle of the permanence of continents and ocean basins.'" Continents were not at all regarded as static then; they bobbed up and down obediently as needful, but certainly they did not wander sideways. By now we have drilled long cores from the ocean bed everywhere, to find in all those thick submarine sediments no fossils older than one fourth of the age of the abundant early fossils in continental rocks: the entire ocean floor is geologically new. Plate tectonics rules the science of geology, its theme the spreading of the sea floor out from midocean ridges until the thin new crust is tucked deep under the shorelines of a dozen shifting plates.

The tale is familiar by now to every geoinformed reader, and the colorful world map, nearly five feet by three, documents the now triumphant theory. A crowd of black dots flows along ocean ridges, through rifts, pooling at coasts and island arcs, 140,000 computer-marked epicenters of archived earthquakes, graded by magnitude. Some 1,500 red triangles, again graded by age of activity, indicate volcanoes that have been active in the past 10,000 years. Arrows give the direction and rate of drift of every rigid plate. This is an eloquent history of unending local catastrophes, disclosed as an orderly if complex dynamics of the whole. The map, on

Mercator projection, lacks the polar zones, but it is impressive for its shadowed relief shown over land and sea floor alike. Color-coded tints give depths and elevations.

The base map is the revised version of a 1985 physiographic map made by the National Oceanic and Atmospheric Administration's National Geophysical Data Center, without borders or cities, bearing only the names of plates. The unmatched richness of digital data here (the base map stands on a grid of four or five million elevations) has led to some deficiencies of appearance, but the authors have so disarmingly included a page of self-criticism of this, their first edition, that only ungenerous readers will be put off. Of course, an improved edition will come. But right now the Hawaiian undersea track visible all the way to Kamchatka, the dense Yellowstone earthquake zones, the press of India into the Eurasian plate to thrust up the Himalayas make it a bargain find for any student of the earth who has wall space.

Only the preface of Albritton's book says much about plate tectonics. His smoothly written and tellingly illustrated text is a cheerful, open account of the history of certain ideas of geologic change. The 13 informative chapters divide into three plates, so to say. The first carries a reader easily from Gilgamesh and Utnapishtim's square ark through the centuries to the end of the period during which catastrophism was in some sense obligatory because of the hasty scriptural time scale. Steno's Maltese finds of shark's teeth and Hooke's ammonites are shown to us, along with those farseeing authors' understanding that fossils were real relics of past life and geologic change.

We read of theories of flood and fire but then come on the wonderful clarity of John Playfair, writing about the slow, organic growth of river valleys. Lyell's field-born gradualism and Darwin's slow selection were hard-pressed no longer by the biblical time scale (which they could evade) but rather by Lord Kelvin's premature geophysical cooling estimates. It was radioactive energy that first gave the geologists world enough, and time.

A few chapters review the next sudden change that fell on the geologists, now happily freed for the continuity they had lacked but always sought. That was the explosion craters. Albritton himself was a pioneer of this understanding, arguing more than 50 years ago that giant meteor impact was real. A sample of half a dozen big

circular cryptocraters is sifted for the reality of the sudden impact: most are past dispute, although possibly not the very largest.

Of course, the next narrative begins in 1980 with the asteroid that slew the dinosaurs, revealed by the iridium content of a half-inch layer of clay from Italy. Five chapters review what has happened since, an explosion of claims pro and con: mass extinctions by asteroids, comets, radiation, bad air, bad seawater, bad rain, darkness. . . . Plainly there is a contagious fever, a "dinosaurmania," that affects even some sober scientists. Did the big reptiles die with a bang or a whimper? Was there a sudden fall or almost as sudden a megavolcanic push? Are there repeated extinctions over the history of plant and animal life? Are they rhythmic? What about soot, shocked granules, microspherules, tektites? The author is agreeable and cool; he takes no sides but presents the arguments of all comers.

"Whatever the outcome of the debate, nobody could say it was dull." There is plenty of evidence to show that something unusual happened at the famous K-T boundary, something outside the evolutionary chain, something that affected the entire world environment. How sudden the change was, whether it came from above or from below (or perhaps from both), whether it was decisive in mass extinction or only a contributor—all remains unclear. The tale is told here up to 1988, tangled and fascinating. A fairer and simpler account you will not find, but don't expect a decision. This reader is most impressed by the four spikes of iridium enrichment reported in 1988 from the original Gubbio boundary clay: not just one sudden bang. A long deep-Pacific core shows only that one iridium-enrichment epoch within the layers of 30 million years; frequent events are unlikely.

We have gained important ground even in our indecisiveness. Changes can plainly be fast or slow; we know that now. The old war of method between dominant catastrophe and the application of present experience to the entire past is now best seen as a regional conflict. "We don't need any more doctrinaire labels." Even the assumption that the laws of energy and matter remain unchanged is now a working hypothesis, tentative but tested to high accuracy over the past billion years or so. The historical sciences are able to set themselves an "ambitious end . . . a unified theory for the unfolding of the universe."

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at \$1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, Utah 84104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of \$1.50 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/89. \$1.50 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the authors' risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 3187, Harlan, IA 51593. Telephone inquiries: 1-800-333-1199, U.S. only; other 515-247-7631/32. The date of the last issue on subscriptions appears on each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

State and ZIP

ESSAY

Ecological diplomacy: an agenda for 1990



by Richard Elliot Benedick

"I am an environmentalist," candidate George Bush declared in 1988. The statement could prove to be as relevant for us as John F. Kennedy's defiant "Ich bin ein Berliner!" at the Berlin Wall was for an earlier generation. For the dangers of East-West confrontation, although far from moribund, are increasingly overshadowed by more subtle threats to the security of humanity. A new set of global environmental dangers demands a new global diplomacy.

Such issues as climate change, depletion of the ozone layer, pollution of oceans and fresh waters, loss of tropical forests, massive extinction of species, acid rain, toxic chemicals and hazardous waste are moving to the top of the world's agendas. Indeed, a nuclear-weapons exchange is arguably more remote than potential ecological collapse resulting from the gradual cumulative impact of hundreds of millions of independent decisions made every day—from applying a chlorofluorocarbon-propelled hair spray to building roads through a forest.

The economic summit of the Group of Seven major industrial democracies held in Paris last summer finally accorded the environment the attention it deserves: one can no longer speak of economic growth without considering the ecological costs. But environmentally responsible policies carry short-term price tags—and these were not addressed at Paris. The U.S. will host the 1990 summit, offering President Bush an opportunity to transform the rhetoric of Paris into meaningful actions. As the world's leading polluter, consumer of energy and generator of wastes, the U.S. bears a special responsibility for assuming leadership.

Dominating the Washington summit must be the issue of finding ways of helping and encouraging the Third World to join in a global effort to preserve the planet. The West has become wealthy even as its industrial, energy and agricultural pol-

icies and consumption habits have overburdened the fragile natural balances on which all life depends. If the coming billions in the poorer countries do no more than try to emulate these patterns, the world's future looks grim. Put quite simply, the cooperation of developing countries will be indispensable.

More than 90 percent of population growth will occur in developing nations. China and India alone account for some two billion people, nearly 40 percent of the world's total. If China, with one third of the world's proven coal reserves, were to exploit its supply fully to fuel economic growth, any efforts by industrialized nations to slow down global warming would be negated. Destruction of tropical forests in Brazil, Indonesia and elsewhere not only aggravates the greenhouse effect but also is leading to unprecedented loss of the planet's biological diversity.

Yet the Third World worries that the new wave of environmental consciousness in Europe and the U.S. means the poor countries will be exhorted to forgo Western-style growth for the greater good of the planet. For example, China and India have made it clear that they will not abandon ozone-destroying chlorofluorocarbons for their refrigeration needs if they must as a consequence buy more costly technologies from the West.

The leaders of the richer nations may be beginning to understand that protecting the global environment is inextricably linked with eliminating poverty in the Third World. As President Bush has stated, "Successful economic development and environmental protection go hand in hand. You cannot have one without the other."

The 1990 economic summit must therefore consider action to confront the interlocked issues of poverty and the environment. This could involve creative initiatives in debt relief (including debt-for-nature swaps, as pioneered by the World Wildlife Fund); subsidized transfer of new energy and other technologies; and expanded and more carefully targeted technical, scientific and financial assistance. The World Bank, with its enormous resources, must be pushed by the Seven and its other major stockholders toward more aggressive promotion of new approaches to energy conservation, sustainable agricultural practices, industrialization, urban planning, reforestation, watershed management, education—and contraceptive research and family planning.

The summit should also strength-

en the United Nations Environment Program, a woefully underfunded and thinly staffed agency that showed its potential by leading the historic negotiations for the Montreal Protocol on protecting the ozone layer. The action-oriented UNEP merits much more support, and the financial implications are minimal: it currently spends less than \$40 million annually.


The summit might well promote specific reforms in national and corporate accounting practices, which currently discourage environmental protection by failing to reflect the real costs to society of pollution and thoughtless short-term exploitation of such natural resources as forests. New economic guidelines are desperately needed to help planners and investors make decisions affecting the security of future generations.

Finally, the Group of Seven countries—by far the major sources of greenhouse gases—should agree to take preemptive steps aimed at mitigating global climate changes. Specified reductions in carbon dioxide emissions, policies to promote greater energy conservation and efficiency, serious attention to expanding nonfossil energy sources, a phaseout of chlorofluorocarbons, and reforestation—together these steps could buy precious time by slowing down greenhouse warming. Such measures would enhance Western credibility with other governments as negotiations begin to involve all nations in sharing the responsibility.

The 1990 summit provides the occasion for leadership by President Bush in an area of vital concern. Scientific uncertainties, technological limitations, entrenched economic interests, wasteful life-styles and political timidity all represent formidable obstacles to action. And yet there is a deep well of public support for an environmental presidency: a 1988 poll revealed that the proportion of Americans who want environmental improvements *regardless of cost* had grown from 45 percent in 1981 to 80 percent.

The world will be watching the participants in the 1990 economic summit for signs of the necessary wisdom and courage.

RICHARD ELLIOT BENEDICK was chief U.S. negotiator of the 1987 Montreal treaty on protecting the ozone layer. Author of *The Ozone Protocol: A New Global Diplomacy*, Ambassador Benedick is now on detail from the State Department as senior fellow of the Conservation Foundation/World Wildlife Fund.



Coats? Umbrellas?
Scarves? You laugh your way
down to the beach.
You've had the last laugh
over winter.
You're in the Canary
Islands.

On a cold winter's day.

You're a happy fugitive.

You've just escaped from the
cold, grey skies and
monotonous rain.

Now, as you dreamily wiggle
your toes in warm sand on the
beach at the Canary Islands,
you toast your escape with a
sip of refreshing fruit juice.

The hours drift by, giving you
time for everything. Sunbathing,
taking refreshing dips in the sea,
going for long walks, savouring
delicious food, capturing
unforgettable moments on film,
enjoying nature, swimming,
windsurfing, snorkelling.

In the evening, as you're
getting ready to go out to one
of many restaurants with a
dance-floor and seaviews, the
moon reflects a new you. No
more dark circles under those
tired eyes, the pale tone of your
skin replaced by a warm,
tropical tan.

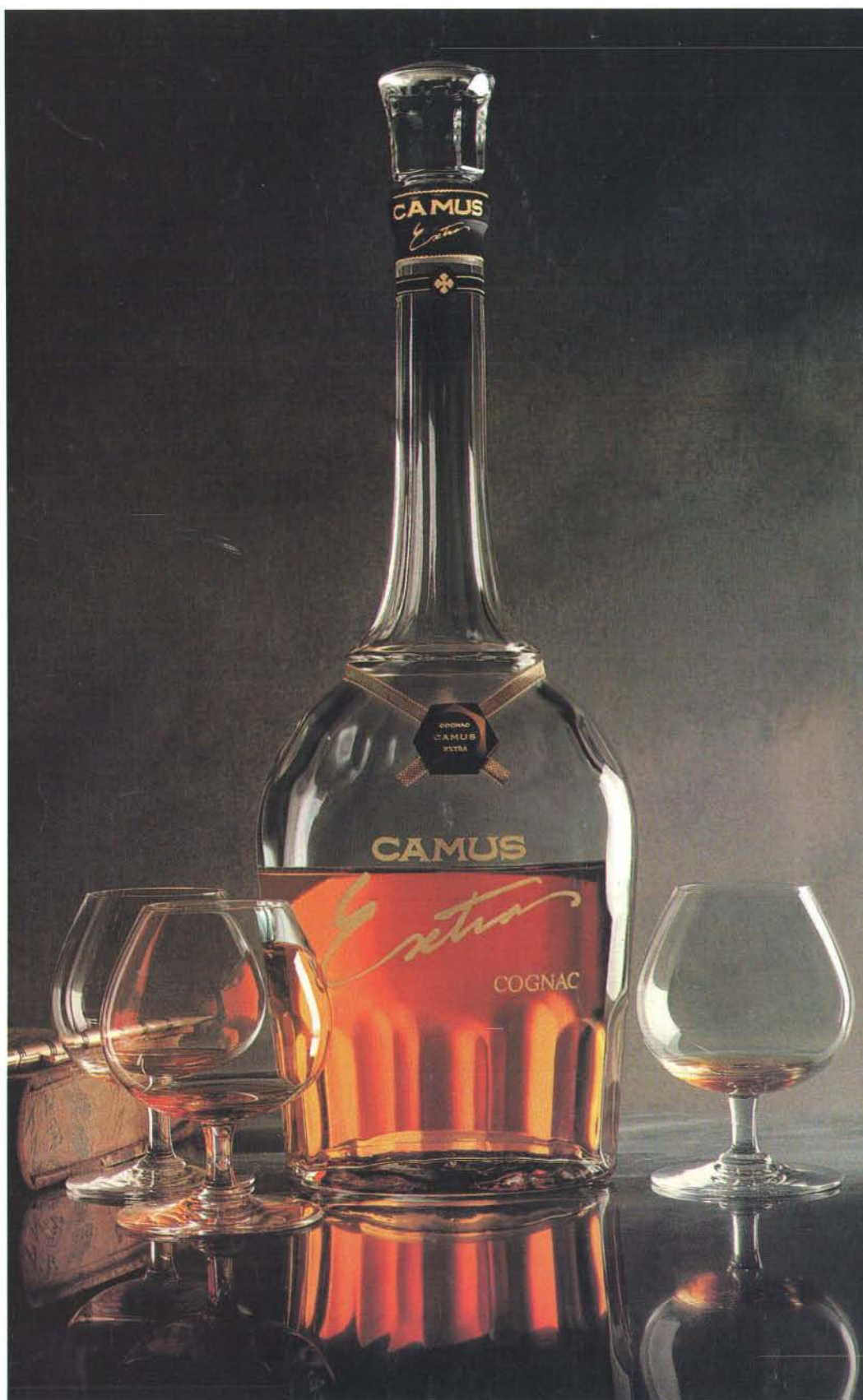
Come visit the Canary Islands
and enjoy its subtropical
climate, the "Costa del Sol", the
Balearic Islands and the
Mediterranean Coast.

Here it's summertime even in
winter.

Spain. Everything
under the sun.



Clearly the judges had no difficulty in voting Camus the best cognac in the world.



In 1984, we at Camus
decided for the first time
to enter our
XO Cognac in the
International Wine and
Spirits Competition.



Camus XO
was deliberated upon
by a collection of
the most highly-qualified
palates in the world,
who duly pronounced
the Camus XO
a worthy winner of the
gold medal.

In 1987, we entered again,
this time with
Camus Extra.



Not surprisingly it, too,
won the gold medal,
leaving Camus with the
enviable record
of two entries and two
gold medals.

Incidentally, no gold
award was given in 1988.

Coincidentally,
Camus did not enter
that year.

CAMUS
COGNAC, FRANCE